Figure 3-1: Actual use of the DEC-2065 for the Month of November, 1988

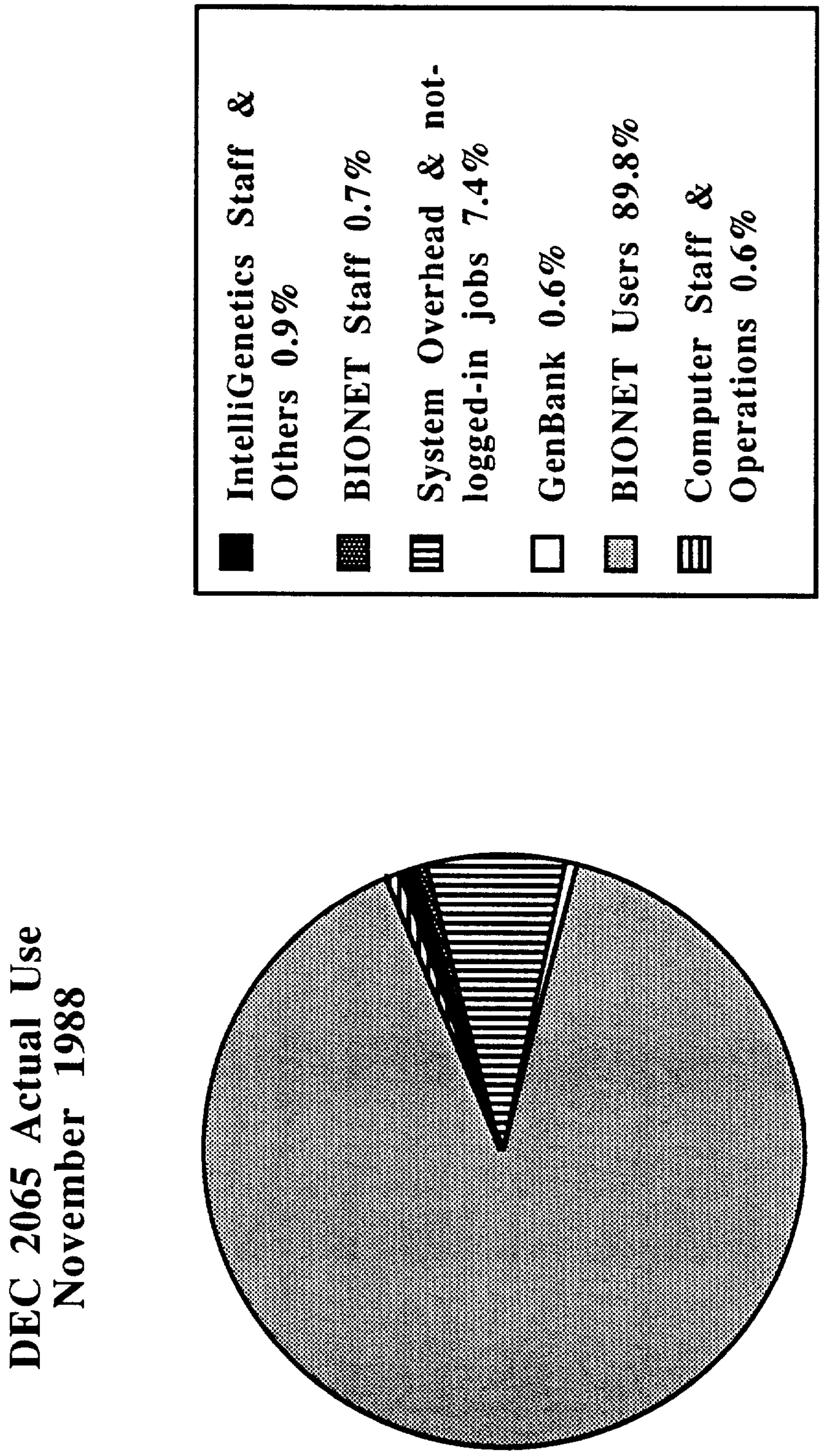
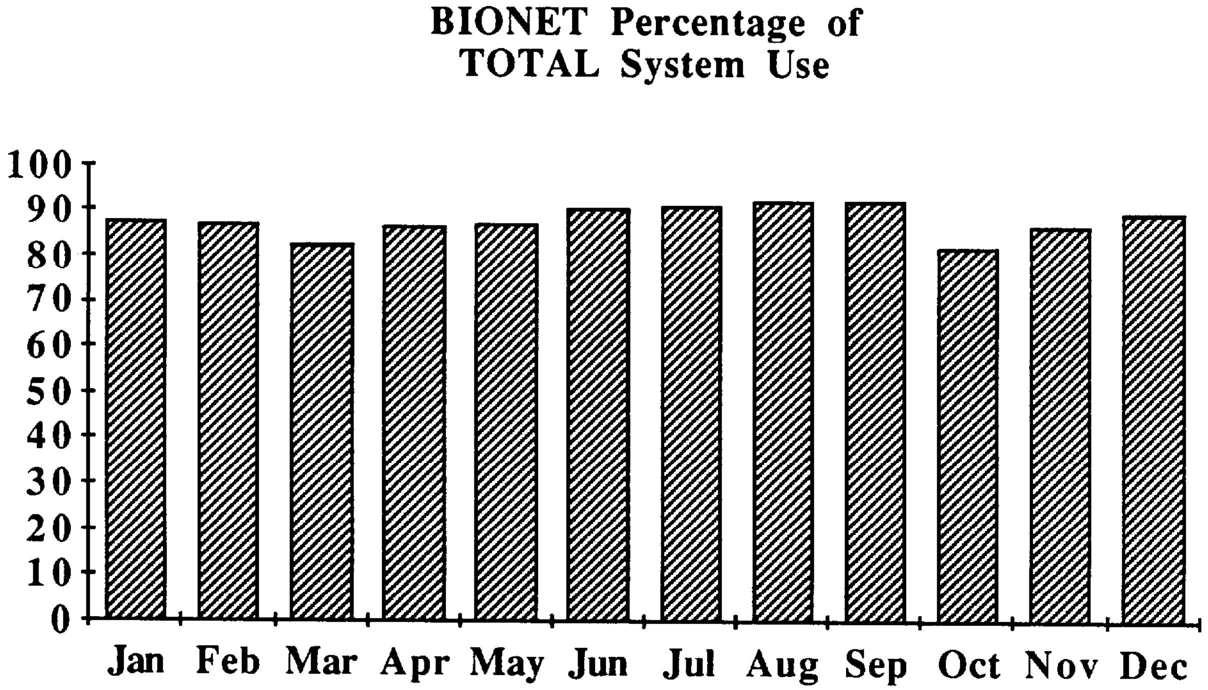


Figure 3-2: BIONET'S Percentage of Total System Use, 12/87 - 11/88



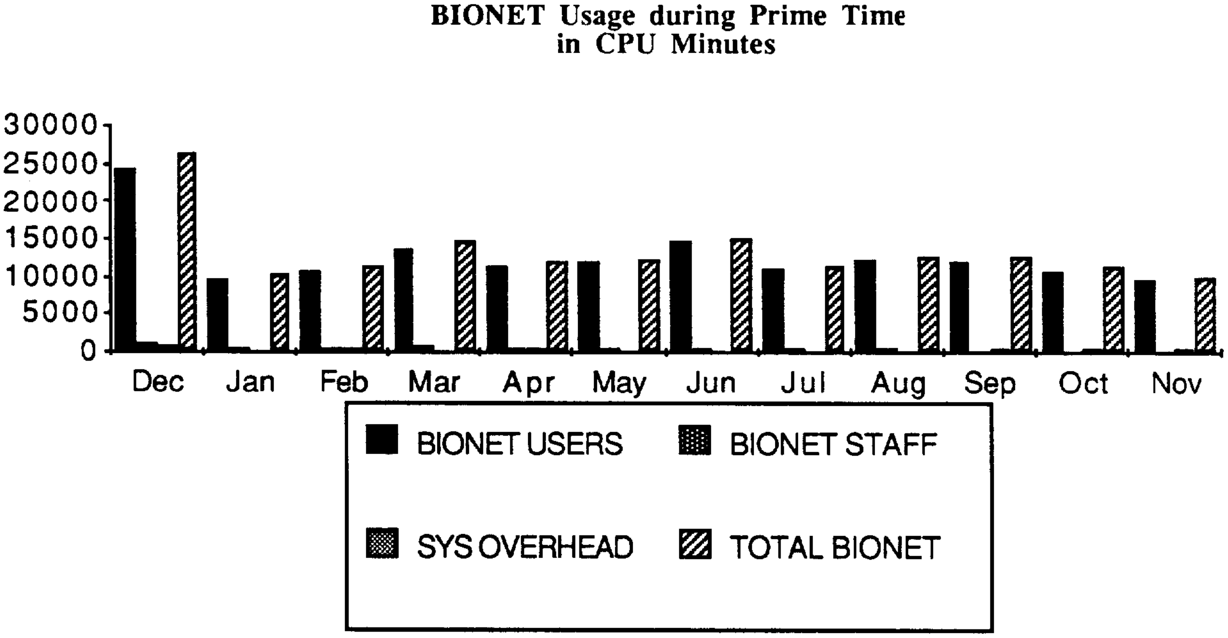
Tabel III-3: BIONET Prime Time CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	24588.20	1076.90	782.60	26447.70
Jan	9840.70	414.00	169.80	10424.50
Feb	10605.80	372.90	335.25	11313.95
Mar	14010.30	649.90	251.45	14911.65
Apr	11521.40	331.10	298.70	12151.20
May	11980.10	391.50	153.45	12525.05
Jun	14747.70	338.70	229.30	15315.70
Jul	10960.00	312.50	258.00	11530.50
Aug	12371.60	288.00	167.40	12827.00
Sep	12261.40	257.90	440.05	12959.35
Oct	10897.40	231.20	322.25	11450.85
Nov	9625.80	176.90	295.45	10098.15
Total	153410.40	4841.50	3703.70	161955.60

Tabel III-4: BIONET Prime Time Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	7798.00	2642.90	4134.35	14575.25
Jan	2718.80	952.80	1782.45	5454.05
Feb	3658.20	867.50	1728.60	6254.30
Mar	5642.74	1257.30	2112.30	9012.34
Apr	4654.20	978.80	1701.85	7334.85
May	4391.40	957.70	1652.45	7001.55
Jun	6065.20	992.00	2040.55	9097.75
Jul	5265.30	668.20	1569.30	7502.80
Aug	5629.10	955.40	1639.50	8224.00
Sep	5168.20	866.30	1928.90	7963.40
Oct	4101.40	773.80	1629.40	6504.60
Nov	3585.20	726.50	1624.30	5936.00
Total	58677.74	12639.20	23543.95	94860.89

Figure III-3: BIONET's Prime Time Use of the DEC-2065 12/87-11/88



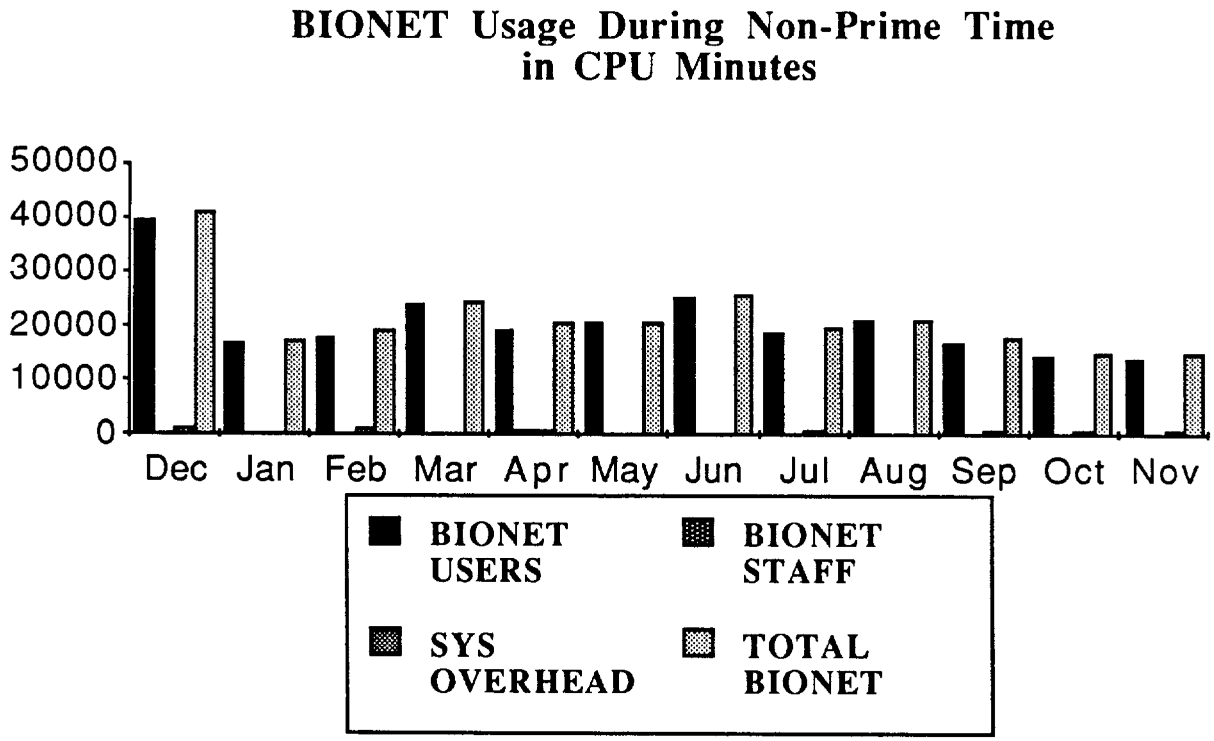
Tabel III-5: BIONET Non-Prime Time CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	39571.00	322.10	1287.75	41180.85
Jan	16832.60	117.00	278.75	17228.35
Feb	17766.40	90.70	1512.10	19369.20
Mar	24050.10	482.30	235.65	24768.05
Apr	19409.10	645.70	759.55	20814.35
May	20637.40	67.90	244.65	20949.95
Jun	25492.50	77.30	267.00	25836.80
Jul	18816.00	175.60	605.25	19596.85
Aug	21007.50	20.30	177.50	21205.30
Sep	17103.50	26.70	742.45	17872.65
Oct	14466.40	16.10	809.05	15291.55
Nov	14094.70	19.20	748.70	14862.60
Total	249247.20	2060.90	7668.40	258976.50

Tabel III-6: BIONET Non-Prime Time Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	5439.10	450.70	5643.70	11533.50
Jan	1690.10	156.10	2640.40	4486.60
Feb	2557.80	117.60	2646.85	5322.25
Mar	4395.60	207.00	3146.75	7749.35
Apr	3170.90	199.70	2597.45	5968.05
May	3258.80	108.00	2579.00	5945.80
Jun	4045.40	87.60	3150.10	7283.10
Jul	4053.40	71.20	2248.00	6372.60
Aug	4743.40	67.70	2549.05	7360.15
Sep	3250.10	105.70	2887.10	6242.90
Oct	2553.00	33.40	2560.30	5146.70
Nov	2194.80	58.50	2524.55	4777.85
Total	41352.40	1663.20	35173.25	78188.85

Figure III-4: Bionet's Non-Prime Time Use of the Dec-2065, 12/87-11/88



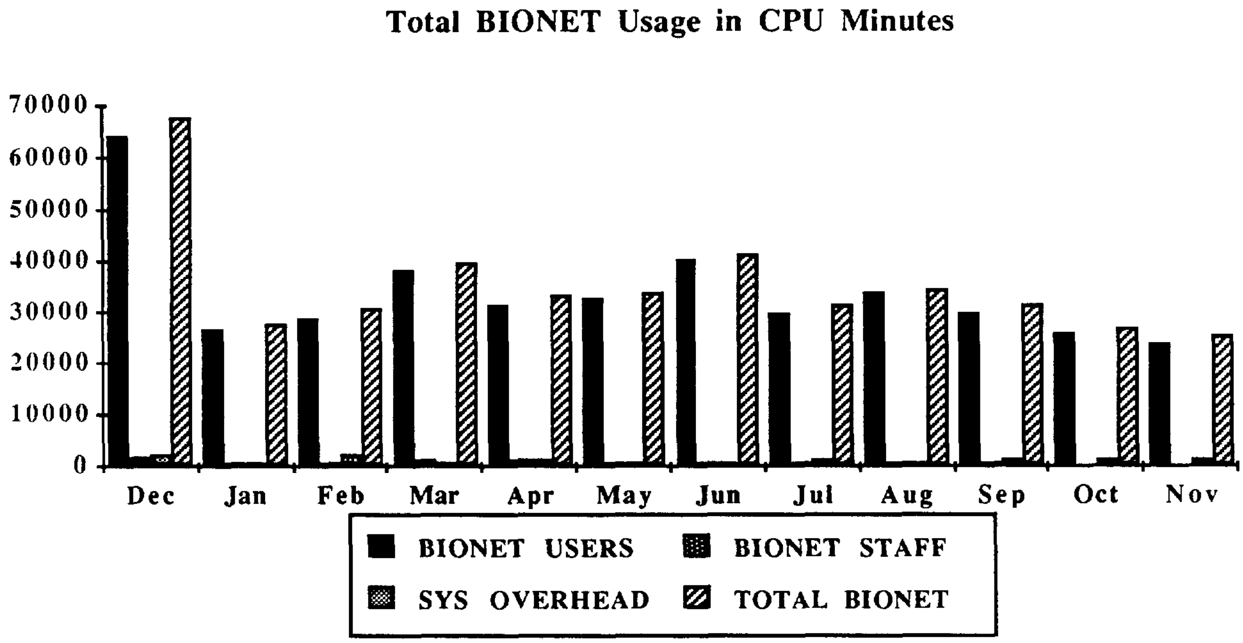
Tabel III-7: BIONET Total CPU Minutes

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	64159.20	1399.00	2070.35	67628.55
Jan	26673.30	531.00	448.55	27652.85
Feb	28372.20	463.60	1847.35	30683.15
Mar	38060.40	1132.20	487.10	39679.70
Apr	30930.50	976.80	1058.25	32965.55
May	32617.50	459.40	398.10	33475.00
Jun	40240.20	416.00	496.30	41152.50
Jul	29776.00	488.10	863.25	31127.35
Aug	33379.10	308.30	344.90	34032.30
Sep	29364.90	284.60	1182.50	30832.00
Oct	25363.80	247.30	1131.30	26742.40
Nov	23720.50	196.10	1044.15	24960.75
Total	402657.60	6902.40	11372.10	420932.10

Tabel III-8: BIONET Total Connect Hours

	BIONET Users Except Staff	BIONET staff	BCRG Plus System Overhead	Total BIONET Use
Dec	13237.10	3093.60	9778.05	26108.75
Jan	4408.90	1108.90	4422.85	9940.65
Feb	6216.00	985.10	4375.45	11576.55
Mar	10038.34	1464.30	5259.05	16761.69
Apr	7825.10	1178.50	4299.30	13302.90
May	7650.20	1065.70	4231.45	12947.35
Jun	10110.60	1079.60	5190.65	16380.85
Jul	9318.70	739.40	3817.30	13875.40
Aug	10372.50	1023.10	4188.55	15584.15
Sep	8418.30	972.00	4816.00	14206.30
Oct	6654.40	807.20	4189.70	11651.30
Nov	5780.00	785.00	4148.85	10713.85
Total	100030.14	14302.40	58717.20	173049.74

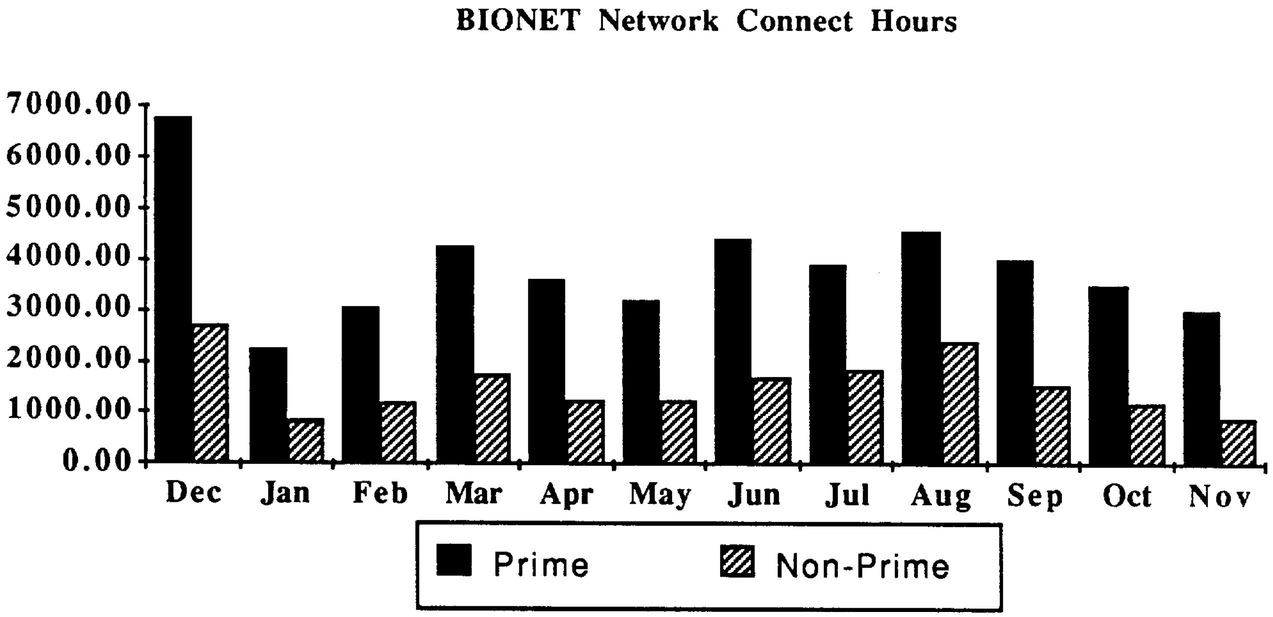
Figure III-5: Bionet's Total Use of the Dec-2065 12/87-11/88



Tabel III-9: BIONET Network Usage Connect Hours

	Prime	Non-Prime	Total
Dec	6779.69	2717.94	9497.63
Jan	2246.27	857.43	3103.70
Feb	3047.92	1203.13	4251.05
Mar	4279.67	1774.58	6054.25
Apr	3620.61	1277.32	4897.93
May	3210.61	1259.72	4470.33
Jun	4421.23	1724.26	6145.49
Jul	3928.71	1860.19	5788.90
Aug	4573.06	2395.25	6968.31
Sep	4009.63	1543.09	5552.72
Oct	3516.30	1228.15	4744.45
Nov	3019.07	902.89	3921.96
Total	46652.77	18743.95	65396.72

Figure III-9: Total Telenet and Compuserve Network Use, 12/87-11/88



3.1.5.4 Computer Software - Core Library

There was a new release (version 5.2) in August 1988 of the IntelliGenetics software suite that makes up the Core Software Library. This software is made available to the BIONET community immediately upon its formal release. Version 5.2 included the addition of new functionality to many of the existing programs and also included a completely new program, FINDSEQ, that simplified the retrieval of sequence data from the GenBank and PIR databases and also allowed for the performance of Boolean searches. FINDSEQ was considerably more user-friendly and powerful than the previously used FIND command, but was designed to handle searches that did not require the power of the IntelliGenetic's QUEST database pattern searching program. The next release of the IntelliGenetics software is scheduled for early 1989 and will be made available on BIONET. It will feature significant new improvements in speed and sensitivity for database similarity searches.

3.1.5.5 Computer Software - System Library

During the course of the year most new systems software was added to the new Sun network. This was documented above in the section entitled **Development work on the new Sun central computing facility**.

3.1.5.6 Computer Software - Contributed Library

Software contributed to BIONET is placed in the <CONTRIBUTED> directory on the DEC-2065, to which only the BIONET community has access. Recently BIONET has also implemented an "anonymous FTP" capability on the new Sun 3/280 file server. This allows users on other Internet computers to retrieve software from BIONET over the network through the use of the FTP or File Transfer Protocol. This method of retrieval is significantly faster than downloading files from BIONET to a remote PC and does not run up communications charges.

Major software packages produced by BIONET collaborators and implemented on BIONET with the aid of our staff have been summarized under *Collaborative Research*, above. Refer also to the software lending library catalog in *Appendix IV*.

3.1.5.7 Database Library

BIONET provides its users with a large number of different databases in support of molecular biology and molecular genetic research, the most popular being the Roberts' restriction enzyme database and the GenBank nucleic acid sequence database. We provide database updates in a timely manner to the community. Our sequence database releases in IntelliGenetics' file format have usually been 2-3 weeks after obtaining the tapes from NIH GenBank or the EMBL. The original database format files are made available on the DEC-2065 usually the same evening of the day on which the tapes arrive.

BIONET databases were discussed in detail in previous Annual Reports. New additions and changes to the databases on BIONET during this year are described above under *Data Contributors*. Here we document the use of the major databases on BIONET.

The DNA and protein sequence databases are used by BIONET scientists as a source of sequence data and for searching. Two major types of searches are performed. The main usage occurs when BIONET users search the database for similarities or homologies with sequences that they have determined. The second type of search involves use of the QUEST program to find interesting consensus sequences that are known to have functional importance. This year our statistics on database searching have been dramatically impacted by the introduction of the FASTA-MAIL program. Since accounting software is not yet available on the Sun system we cannot provide accurate data for the number of individual GenBank, PIR, or SWISS-PROT database searches performed by users. Statistics can, however, be provided for the overall total number of searches. We know that the FASTA-MAIL interface on the DEC is utilized 950 times per month or 32 times per day on average. FASTA-MAIL is used to search each of the three databases just mentioned. BIONET has successfully promoted the use of FASTA-MAIL over the use of other database searching software to reduce the load on the DEC 2065 computer. In addition the IntelliGenetics' QUEST and IFIND programs are used 346 times per month for database searches.

Restriction Enzyme Database - Thanks to Dr. Richard Roberts, the chairman of the BIONET National Advisory Committee, we have established one of the most up-to-date lists of restriction enzymes available. Dr. Roberts maintains a restriction enzyme registry and distributes his updated lists in an electronic message to BIONET. BIONET incorporates these new lists into its program essentially immediately. These lists are used within the core program SEQ and PEP and are referenced 459 times per month.

VectorBank^{tm}. - Vectorbank is a collection of sequence data and restriction maps of important cloning vectors, viruses and phages that is maintained by the IntelliGenetics staff. This database is used by the CLONER program for manipulating restriction maps and simulating DNA cloning experiments. Release 4.2 of VectorBank in May saw the revision of the restriction map data files for all 146 vectors in the database. All single and double cutter sites using prototype enzymes from the latest restriction enzyme database were included in the map data. Individual Vectorbank sequence files are accessed up to 25 times per month.

KeyBank^{tm} - KeyBank is a data bank of nucleic acid and protein consensus patterns collected from the literature by IntelliGenetics, Inc. KeyBank is produced in a format suitable for use with the QUEST pattern searching program of the IntelliGenetics Suite. Some of the patterns available include binding sites, active sites, allosteric sites, phophorylation sites, cleavage sites, cap sites, chromosome modification sites, polymerase binding sites, insertion sites, regulatory sites, methylation sites, replication origins, satellite DNA, Z DNA, and zinc finger regions. The latest release (rel. 3.0, October 1988) has 1237 new patterns taken from more than 900 references. This part of KeyBank occupies more than 690,000 bytes of memory. In addition, KeyBank also has files of codon and restriction enzyme site patterns. The contents of the key files in the main part of KeyBank can be easily searched by using the many indices supplied with the database, such as the organism, keyword, or citation indices. Individual database files in KeyBank are utilized up to 10 times per month.

3.2 Highlights

The sections above describe in detail our accomplishments in the several components of the BIONET Resource. Considering that a very signficant fraction of staff time during our fifth year was involved with planning, grant writing, and other obligations for our renewal, substantial progress was still made on the Resource. Here, in brief, are some of the most notable:

- The new computer network donated by Sun Microsystems is being prepared for direct access by the user community by the BIONET systems staff (Mr. Rob Liebschutz and

Mr. Eliot Lear). As of August 1988 BIONET released the FASTA-MAIL program. This provided our users on the DEC with electronic mail access to high-speed database searches on our Sun 3/280 computer. Database search times were reduced from hours to tens of minutes or less, and the average mid-day user load on the DEC 2065 **dropped by a factor of about three to five** since it was no longer being used for these compute-intensive tasks. FASTA-MAIL used the FASTA program obtained from Dr. William Pearson, and the mail server portion was developed by Mr. Liebschutz, Mr. Lear, and Mr. Spencer Yeh.

- The BIONET user community continued its vigorous growth, up 31% over last year for a total of 867 laboratories on the system. On the DEC 2065 total cpu usage increased by 65% and total connect hours by 36% as compared to last year.
- Dr. Jerzy Jurka, the BIONET Scientist, published important work in the area of repetitive DNA sequence analysis. In conjunction with this research, new functionality has been added to the Multiple Aligned Sequence Editor (MASE) by the BIONET applications programmer, Mr. Liang Jen Horng. This editor was originally developed by Dr. Jurka and Donald Faulkner at Dana Farber's Molecular Biology Computer Research Resource and then extended at BIONET over the past year. Work on the editor has also involved collaborators from the machine learning group at the University of California at Santa Cruz.
- Dr. Sunil Maulik has continued work on the RICH program which performs database searches for sequences of defined percent composition. Dr. Maulik has obtained some interesting preliminary results with the program. He has also been involved in developing a new Hypercard^{tm}-based user interface for BIONET.
- The electronic communications network was significantly enhanced. The efforts of Dr. David Kristofferson led to the formation of the international BIOSCI bulletin board network. Besides BIONET in the U.S., other major BIOSCI distribution sites are situated in England, Ireland, Sweden, and Finland. Recipients of the bulletin boards from these sites are located in all parts of the globe. The bulletin boards are available to users on the ARPANET, BITNET, EARN, Usenet, NSFnet, and JANET. Users in any particular location need only post or receive messages from their closest site. Any posting at any center is automatically forwarded by the central BIOSCI sites to all other participants on the above-listed networks.
- Finally, the research conducted by BIONET's 867 laboratory groups was made significantly easier by a total revision of the BIONET documentation and the production of a new User Manual. This involved major efforts by BIONET staffer's Ms. Vickie Johncox, Mr. Spencer Yeh, and Ms. Kathryn Berg. The documentation was sent free of charge to all users on the system this past summer.

3.3 Administrative Changes

The following have been the administrative changes within BIONET during the past year. These have come about for reasons ranging from personnel shifts to additions and resignations. None of these changes has had a negative impact on the Resource itself, in particular, its "appearance" and availability to the community.

- Mr. Liang Jen Horng was hired in May 1988 as the new BIONET Applications Programmer to assist Dr. Jurka in his research efforts and to work on the BIONET contributed software. Mr. Horng has a M.S. in Computer Science from San Jose State University.
- Mr. Eliot Lear was hired in June by the Computer Facilities group at IntelliGenetics as an additional system programmer. Mr. Lear has a B.S. in Computer Science from Rutgers and has run a state-wide network of Sun Microsystems file servers and

workstations. Since joining IntelliGenetics he has devoted a large fraction of his time to bringing up the new Sun system for BIONET.

- Ms. Vickie Johncox, formerly a BIONET Scientific Consultant, left BIONET in June for another job as head of IntelliGenetics' training group. BIONET hired Dr. Karen Davis, formerly at the University of California, Santa Cruz, as her replacement. Dr. Davis started at the end of June. No disruption in the service occurred as Ms. Johncox's duties were handled smoothly by Mr. Spencer Yeh and Drs. Maulik and Kristofferson.
- In November, Ms. Kathryn Berg resigned as BIONET Administrator to pursue work elsewhere. Before her departure BIONET hired Ms. Cindy Eppard from IntelliGenetics to replace Ms. Berg. Ms. Eppard learned her new responsibilities rapidly, and there was no interuption in the establishment and processing of BIONET user accounts.

Despite the demands of their jobs, increased once again by a 31% growth in the size of the user community this year, the BIONET staff remains a highly dedicated group of individuals who are interested in their work and in promoting the goals of the resource. BIONET has received many testimonials to the quality of its staff and the support that they provide. Copies of a few of these testimonials are included in *Appendix VIII.*

3.4 Resource Advisory Committee and Allocation of Resources

The membership of BIONET's National Advisory Committee has not changed this past year. The NAC consists of:

- Dr. Richard J. Roberts, Ph.D., **Chairman**, Senior Staff Investigator, Molecular Biology, Cold Spring Harbor Laboratory
- Professor John Abelson, Ph.D., Department of Biology, California Institute of Technology
- Professor Alan Maxam, Ph.D., Dana Farber Cancer Institute, Harvard Medical School, Harvard University.
- Thomas Rindfleisch, M.S., Director, Knowledge Systems Laboratory, Department of Computer Science, Stanford University.
- Professor Irwin Kuntz, Ph.D., Department of Pharmaceutical Chemistry, University of California, San Francisco.
- Professor Charles Yanofsky, Ph.D., Department of Biological Sciences, Stanford University.
- Professor Eric Lander, Ph.D., Harvard Business School
- Professor Joshua Lederberg, M.D., Ph.D., President, The Rockefeller University.

The last regularly scheduled meeting occurred on November 13, 1987 in Mountain View. Due to the time occupied by preparations for the BIONET renewal the next NAC meeting has been postponed until the first part of 1989. When issues have arose during this last year we have consulted members of the NAC by phone, especially our chairman, Dr. Roberts, and Mr. Thomas Rindfleisch at Stanford.

Allocation of Resources. The Committee agrees with our methods for allocating the Resource. The DEC-2065 computer uses its windfall scheduler to allocate cpu time to the various categories of users and overhead, as described under *Resource Facilities.* The cpu time is distributed on a first-come, first-served basis. This method has been very successful. Considerably more than 50% of CPU time (BIONET's original allocation) has been delivered to BIONET scientists (see *Resource*

Facilities, above). BIONET is now routinely using almost 90% of the DEC 2065 plus significant resources on the new Sun computer system.

We continue to request that the community not have more than one person per PI group using BIONET at the same time during prime time. The community continues to do an excellent job in complying with this policy.

We continue to allocate additional disk space to PI groups involved in managing large sequencing projects or extensive databases of sequences. We do this on an *ad hoc* basis upon requests by investigators. Our archive and retrieval system is working smoothly for archival storage and prompt retrieval (one to two days) of important files.

3.5 Dissemination of Information of Resource's Capabilities

We discuss two areas related to dissemination of information about the Resource that we have pursued this grant year. The first is interactions with the scientific community through participation at conferences and seminars. The second is use of the electronic mail and bulletin board facilities of the Resource itself to keep scientists worldwide aware of changes and improvements.

3.5.1 Community Interactions and Awareness

We have used two methods this year to inform the community about BIONET and to solicit applications for access to the Resource. The first method has been the presentation of invited seminars and participation at major conferences where we have presented lectures and/or have had booths at exhibitions. These efforts are summarized above under *BIONET Training Program*. At these conferences, we have distributed the standard applications packets to scientists, after demonstrating to them the capabilities of the Resource.

The second method has been through the publication of journal articles which describe BIONET. This year an article detailing new developments at BIONET appeared in vol. 16 no. 5 of *Nucleic Acids Research*. A preprint of this article was in last year's Annual Report. An upcoming article will appear in *Protein Sequence and Data Analysis* and is listed on the BRTP Training Form under *Scientific Subprojects*. A copy of this article is available in *Appendix II*.

3.5.2 Electronic Communications

The electronic communication facilities of BIONET provide another important way to disseminate information about the Resource. In addition, electronic mail and bulletin boards provide a mechanism for scientific and technical interchanges among members of the community. With the dissemination of BIONET bulletins to investigators outside of the Resource (via ARPANET, BITNET and USENET), information about BIONET is being distributed electronically worldwide. Information on the types of electronic mail communications with BIONET was summarized previously in the discussion of the *Collaborative Research* component of the Resource.

3.6 IMPORTANT Suggestions and Comments

For the BIONET staff the main "highlight" of this year was the renewal process of the BIONET grant. We have not commented on this topic in this report because some issues are still under negotiation with the NIH and are not yet for public disclosure. We do, however, have the following general comments.

The regulations of the BRTP program at the Division of Research Resources require a Resource such as BIONET to excel in several different areas: Technological Research, Collaborative Research, Service, Training, and Dissemination. Given the size of the BIONET user community and the demands that it makes on our staff, it has always been a challenge to cover all of these areas adequately. We have excelled in the areas of Service, Training, and Dissemination because, until the addition of Dr. Jurka, that was where our talents lay and the sheer din of the user community would have been deafening had we neglected the Service to devote more of our limited resources to research. We feel that the real justification of this resource has been the quality of the research done **BY OUR HUNDREDS OF USERS** and that any efforts on our behalf would always be dwarfed by comparison.

Nonetheless, DRR regulations make it possible for a committee of ten people **who are not even users of the Resource** to come in and possibly scuttle the entire operation because of deficiencies in just **one** of the five mandated areas. Furthermore it is possible that such an event can occur without **ANY** input from the almost 3000 users of the resource! The fact that such an action could lead to the disruption of research and electronic communications for almost 900 laboratories around the world does not seem to be a perturbing factor in the renewal decision. **This surely must be one of the more amazing possibilities in the history of modern science!**

If this year's report demonstrates anything, it shows that **BIONET works and works successfully**, and that it has continued to make significant improvements over the last several years. We have the largest user community of any similar resource in the world and it continues to grow, even attracting users from locations that have competing facilities. Few, if any, other related organizations can lay claim to the "resourcefulness" that BIONET has shown in successfully pursuing its proposal to Sun Microsystems which resulted in the grant of $150,000 in new hardware.

We suggest that reason finally prevail in this process, and that the NIH recognize, preserve, and expand this valuable resource. DRR regulations should be reviewed and revised so that organizations that are primarily devoted to Service need not excel in every single BRTP category in order to survive.

To date the user community is only dimly aware of the forces impinging upon BIONET. Because of the current negotiations with the NIH it has not been appropriate to involve them. It would be extremely interesting to see their reaction if events lead to the termination of the BIONET Resource, but, if reason prevails, this disasterous event will not come to pass, and the users will continue to invest their energies productively in their research.

I. BIONET Research Publications

Copies of six BIONET research publications and abstracts are included in this section.

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 4775–4778, July 1988
Evolution

A fundamental division in the *Alu* family of repeated sequences

(evolution/*Alu* subfamilies/secondary structure/CpG dinucleotide)

JERZY JURKA*† AND TEMPLE SMITH‡

*Bionet, 700 East El Camino Real, Mountain View, CA 94040; and ‡Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, MA 02115

Communicated by Roy J. Britten, March 10, 1988

ABSTRACT **The *Alu* family of repeated sequences from the human genome contains two distinct subfamilies. This division is based on different base preferences in a number of diagnostic sequence positions. One subfamily of the sequences, referred to as the *Alu*-J subfamily, is very similar to 7SL DNA in these positions. The other subfamily, *Alu*-S, can be divided further into well-defined branches of sequences. These findings revise the previous picture of the *Alu* family and expose their complex evolutionary dynamics. They reveal sequence variations of potential importance for the proliferation of *Alu* repeats and relate them to their structural features. In addition, they open the possibility of using different types of *Alu* sequences as natural markers for studying genetic rearrangements in the genome.**

A typical human *Alu* family member is a sequence ≈300 base pairs long and consists of two similar but not identical subunits, *Alu*-left and *Alu*-right, connected by an adenine-rich linker. Both halves of *Alu* elements are related to the 7SL RNA (1). Although *Alu* sequences are the most abundant among middle repetitive elements in the human genome, their biological role remains unclear (2). In this paper, we report on the presence of at least four different types of *Alu* sequences, which probably originated at different times in the history of primates.

METHODS

A set of 125 complete or nearly complete human *Alu* sequences were extracted from the GenBank DNA sequence data base.§ The list of the GenBank loci used, positions of the extracted sequences, and other specifications are given in the legend of Table 1.¶ The statistical analysis described below is based exclusively on pairwise comparisons of each *Alu* sequence with the consensus sequence (see Fig. 1), using the computer algorithm of Smith and Waterman (3). The overall consensus sequence in Fig. 1 was derived from our data and is slightly different from the one recently published (2). The differences are exclusively within CpG doublets, which are known to be variable in *Alu* repeats (4). Taking pairwise comparisons as a starting point, the multiple alignment of the analyzed set of sequences has been done by hand with a specialized sequence editor (5). To detect sequence positions with different base preferences (diagnostic positions), we used "column-correlation" function incorporated in the sequence editor (5). This function was originally designed to perform automatic searches for compensatory mutations.

RESULTS

During a search for compensatory mutations in the multiply aligned set of 125 *Alu* sequences, we noted an unusually high proportion of correlated base occurrences in at least 15 sequence positions. These positions are referred to as diagnostic positions and are listed in column 1 of Table 1 (the position numbers are the same as in Fig. 1). The observed correlations in the diagnostic positions reflect different base occurrences in different *Alu* subfamilies. It is shown below that the most predominant bases in the 15 diagnostic positions belong to only one of the two basic types of *Alu* sequences present in the analyzed set.

To segregate the most predominant type from the remaining *Alu* sequences, we have used computer alignment (3) of each *Alu* element with the *Alu* consensus from Fig. 1. The average overall similarity between the 125 *Alu* sequences and the *Alu* consensus is 83.88% with a SD of 5.63% (gaps counted as single mismatches). These numbers are slightly different if gaps are excluded from the analysis (see Table 3). Given the overall similarity, we assume that the probability of matching between any *Alu* sequence and its consensus in a randomly chosen aligned position equals 0.83. Any *Alu* sequence similar 40% or less to the consensus sequence in the 15 diagnostic positions has been defined as an *Alu*-J element. The probability of only 6 matches or less in 15 randomly chosen aligned positions can be calculated from the binomial distribution and is <0.001. Following the statistical definition, we have found 31 *Alu*-J sequences in the analyzed set of 125 sequences. The remaining 94 sequences are referred to as *Alu*-S sequences. The 3:1 ratio of S/J *Alu* sequences explains why the overall consensus sequences and *Alu*-S consensus sequence overlap. We have found no sequences matching seven or eight diagnostic consensus positions, which suggests that the distinction between J and S sequence types is quite sharp with few or no intermediate forms. As shown in Table 1, in the diagnostic positions the J subfamily maintains consistently different bases from those in the S subfamily. The difference in base preferences between J and S subfamilies is most evident at positions 94, 204, and 275 (Table 1). For example, G-204 is present in 29 of 31 *Alu*-J sequences and in only 1 of 94 *Alu*-S sequences. Similarly, G-94 and C-275 are powerful diagnostic indicators that can be used for preliminary "by eye" identification of *Alu*-J elements.

As illustrated in Table 1 the most frequent bases in the J subfamily are identical with those in 7SL RNA in 14 of 15 diagnostic positions. Furthermore, the differences between J and S *Alu* elements correlate with differences in the adenine-rich linker connecting the left and right halves of the *Alu* dimer (positions 121–133 of the consensus sequence in Fig. 1; data not shown). The triplet TAC in the middle of the linker is present in ≈80% of *Alu*-S compared to only 20% of the *Alu*-J sequences. It is not certain if the homologous TAC triplet was ever present in many *Alu*-J sequences since their

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed.

§EMBL/GenBank Genetic Sequence Database (1987) GenBank (IntelliGenetics, Mountain View, CA), Tape Release 46.0.

¶The *Alu* sequences used in this study are available on the Bionet computer in the file <jurka>human-alu.seq. The data can also be obtained by electronic mail from jurka@bionet-20.arpa.

Table 1. Diagnostic base differences between major subfamilies of the *Alu* family

Consensus	*Alu*	Frequency of					Base in	Consensus	*Alu*	Frequency of					Base in
position	subfamily	T	C	A	G	(−)	7SL DNA	position	subfamily	T	C	A	G	(−)	7SL DNA
57 (C)	J	6	3	20	0	2	A	163 (A)	J	1	2	3	25	0	G
	S	36	47	9	2	0			S	2	1	55	36	0	
63 (A)	J	2	0	12	16	1	G	194 (A)	J	1	0	6	22	2	G
	S	1	2	87	4	0			S	1	1	91	1	0	
65 (C)	J	30	1	0	0	0	T	204 (A)	J	0	0	1	29	1	G
	S	20	41	1	1	37			S	0	0	89	1	2	
70 (G)	J	2	21	0	7	1	T	208 (G)	J	0	0	20	11	0	A
	S	0	2	3	89	0			S	2	3	29	59	1	
71 (T)	J	5	23	0	2	1	C	220 (T)	J	6	22	1	1	1	C
	S	90	4	0	0	0			S	80	10	0	1	3	
94 (C)	J	0	0	2	29	0	G	233 (A)	J	22	2	4	0	3	T
	S	4	87	1	1	1			S	0	0	90	3	1	
101 (G)	J	2	0	19	10	0	A	275 (T)	J	1	28	1	0	1	C
	S	0	0	9	85	0			S	85	2	1	2	4	
106 (A)	J	0	0	13	18	0	G								
	S	0	0	93	1	0									

Consensus positions are taken from Fig. 1. (−), Alignment gaps. Loci names and 5′ → 3′ positions of *Alu*-J and *Alu*-S sequences are listed below as they appear in GenBank (release 46.0).§ *Alu* repeats complementary to the consensus sequence are listed in 3′ → 5′ order. Positions preceded by b and c indicate b and c branches of *Alu*-S sequences, respectively, as defined in Table 2 and in the text. *Alu*-J: HUMACHRA7(1580–1295); HUMADAG(4907–5201, 24773–24495, 31460–31747); HUMAPOCII(1982–2235); HUMAPOE4(2562–2849); HUMBLYM1(266–560); HUMERPA(1810–2100); HUMFIXG(24172–24465); HUMFOL5(1577–1847); HUMIFNB3(2663–2405, 13213–13489); HUMIL2R8(1209–1486); HUMLDLR(4193–4485); HUMPOMC2(26–340); HUMPOMC6(26–303); HUMRSAOLD(498–790); HUMRSKPA1(24–291); HUMTBB5(2922–2627, 2949–3239, 5611–5885); HUMTHBNB(3593–3883); HUMTPA(7512–7227, 8862–9165, 10801–10513, 16794–17114, 18878–19167, 20944–21250, 22262–22536, 26941–27228); M13121(1141–856). *Alu*-S: HUMA1ATP(4932–5219); HUMADAG(1672–1369, 2357–2642; c: 5606–5893, 8000–7720, 8484–8193, 13452–13741, 15386–15096, 15806–16094, 17224–16933, 18414–18706, 19900–19613, 22527–22812, 25453–25163, 27269–26979; c: 28032–28320); HUMAGG(b: 1391–1106); HUMALBG(3287–3576, 6046–5759); HUMANFA(c: 1340–1621; b: 1630–1919); HUMAPOAI1(3291–3585, 6421–6709); HUMAPOAII(2571–2860); HUMAPOCII(2254–2542); HUMAPOE4(636–352, 2427–2138, 5049–4773); HUMC1A21(347–60); HUMC1A23(330–45); HUMC1AIN1(992–1285); HUMFIXG(7298–7595; c: 31537–31801, 35947–36248); HUMFOL5(1284–989); HUMGAST2(187–477); HUMGHV(2506–2248); HUMHBA4(2060–1773, 4297–4585, 8548–8836); HUMHBBRT(482–190, 1260–1548); HUMIFNB3(4648–4363, 7265–7545; c: 8975–8688); HUMINS2(69–357); HUMLDLIVS(291–8); HUMLDLR(b: 3715–4011); HUMMHDC3B(b: 3712–3424); HUMMHDRB3(b: 2838–3124; c: 4063–4345); HUMMYCRT(c: 3143–2876); HUMNGFB(c: 5259–5544); HUMPOMC(1392–1102, 7099–6803); HUMPOMC1(333–47); HUMRSA1(c: 508–803); HUMRSA27(1–251); HUMRSA16(b: 168–451); HUMRSAB11(1–269); HUMRSAB13(11–295); HUMRSAB19(1–241); HUMRSAB2(1–288); HUMRSAB6(1–256); HUMRSAB8(1–265); HUMRSAP3(b: 897–1186); HUMRSKA1(21–347); HUMSLJT1(568–280); HUMTBB5(3289–3573, 4115–3849; b: 5241–4953; b: 6799–6516); HUMTBBM40(c: 1828–2113); HUMTHBNB(1165–874, 3418–3110); HUMTPA(5960–5671, 6746–6483, 739–1022, 10066–10355, 12986–12700; b: 17170–17455, 21279–21567, 21940–21651, 25619–25905, 26522–26811, 27879–28149; b: 28803–29090, 32922–33210, 34234–34503); HUMUG2PD(c: 546–260, 1685–1396); M11591(b: 1404–1115); M12036(637–362); M12929(592–302).

linkers vary substantially in both their size and the primary sequence.

Following the same approach, the S subfamily of the *Alu* family has been found to contain other types of *Alu* sequences. Unlike the J/S division, the intra-S division is more difficult to define statistically since the number of simultaneous differences between subsets of *Alu*-S sequences appears to be smaller and there is a number of intermediate sequences virtually absent from the J–S junction. Therefore, we first define the most distinct "b" branch of the S subfamily as containing sequences that match 3 or fewer of the 11 diagnostic positions listed in Table 2 and in Fig. 1. There are 12 such elements in the analyzed set of 94 *Alu*-S sequences. The average overall similarity between each analyzed *Alu*-S sequence and the consensus sequence is 86.59 if every gap is counted as a single mismatch (Table 3). Based on this number, we assume the probability of matching the aligned consensus sequence at a randomly chosen position to be 0.86. As calculated from the binomial distribution, the probability of matching 3 or fewer of the randomly chosen aligned positions is 10^{-4}. The probability of matching exactly 4 and 5 positions equals 1.9×10^{-4} and 1.63×10^{-3}, respectively. We have found 5 sequences matching 4, and 6 matching 5 diagnostic positions in the analyzed set of 94 *Alu*-S sequences. These 11 sequences are arbitrarily defined as a "c" branch of the S subfamily. After segregation of the b and c branches, the remainder of the *Alu*-S subfamily is referred to as an "a-branch." Preliminary analysis of 71 *Alu* elements from this branch revealed the presence of 16 sequences containing simultaneously thymine at position 244 and adenine at position 272, as opposed to C-244 and G-272 in the remaining 55 *Alu* sequences. In addition, 14 of the above 16 sequences contain an extra adenine in position 264. This suggests that the *Alu*-a branch may contain at least two different types of *Alu* sequences and it can tentatively be replaced by "d" and "e" branches containing 16 and 55 sequences, respectively.

As illustrated in Table 2 and Fig. 1, the base preferences are quite similar between *Alu*-b and *Alu*-c sequences up to position 88. Further on, *Alu*-c remain similar to *Alu*-a with the exception of guanine at position 163. Therefore, the c branch can be viewed as an intermediate between the a and b branches of the S subfamily. A unique feature of the c sequences may be the presence of adenine at position 74. Of the 125 *Alu* sequences only 8 contain A-74 of which 7 belong to the *Alu*-c branch defined above. The eighth *Alu* sequence containing adenine at position 74 (HUMPOMC1) has all the *Alu*-c features listed in Fig. 1: deletion at 64 and 65, A-78, T-88, and G-163. Therefore, it can also be considered as an *Alu*-c sequence.

Based on the analysis of phylogenetic trees, other authors (6) have recently identified the *Alu*-b branch as a "subfamily of the *Alu* family." The authors have pointed out differences between the *Alu* consensus and the *Alu*-b sequences at positions listed in Table 2 as well as in Fig. 1 and at other less characteristic positions not included in our analysis.

```
           1            15 16           30 31           45
7SL        -GCCGGGCGCGGTGG CGCGTGCCTGTAGTC CCAGCTACT-CGGGAG
Alu-cons   GGCCGGGCGCGGTGG CTCACGCCTGTAATC CCAGC-ACTTTGGGAG

           46           60 61           75 76           90
7SL        GCTGAGGCTGGAGGA TCGCTTGAGTCCAGG AGTTC....CCAGCC
Alu-J                 a      g t    CC
Alu cons   GCCGAGGCGGGCGGA TCACCTGAGGTCAGG AGTTCGAGACCAGCC
Alu-c                          --      (A)   A         T
Alu-b                          --            A         T

           91          105 106         120 121         135
7SL        TGGGCAACATAGCGA GACCCCGTCTCT
Alu-J         G      a     g
Alu cons   TGGCCAACATGGTGA AACCCCGTCTCTACT AAAAATACAAAAATT
Alu-c
Alu-b          T    C

           136         150 151         165 166         180
7SL        -GCCGGGCGCGGTGG CGCGTGCCTGTAGTC CCAGCTACTCGGGAG
Alu-J                                  g
Alu cons   AGCCGGGCGTGGTGG CGCGCGCCTGTAATC CCAGCTACTCGGGAG
Alu-c                                  G
Alu-b                        G         G

           181         195 196         210 211         225
7SL        GCTGAGGCTGGAGGA TCGCTTGAGTCCAGG AGTTCTGGGCTGTAG
Alu-J                   G          G   a            C
Alu cons   GCTGAGGCAGGAGAA TCGCTTGAACCCGGG AGGCGGAGGTTGCAG
Alu-c
Alu-b                       G  R                   C

          226          240 241         255 256         270
7SL       TGCGCCTGTGA....G CCACTGCACTCCAGC CTGGGCAACATAGCG
Alu-J             T
Alu cons  TGAGCC-GAGATCGCG CCACTGCACTCCAGC CTGGGCGACAGAGCG

           271         285
7SL        AGACCCCGTCTCT
Alu-J          C
Alu cons   AGACTCCGTCTCAAA AAAAA
```

FIG. 1. Consensus sequence for 125 *Alu* sequences and the homologous regions of human 7SL DNA. Major and minor characteristic bases for other types of *Alu* sequences are printed in capital and lowercase letters, respectively, and correlate with the analysis in Table 1. Dots indicate sequence regions absent from 7SL DNA but present in the *Alu* family. The remaining 7SL-specific sequences are not shown. Dashes under positions 64 and 65 indicate bases missing in *Alu*-b and *Alu*-c sequences. Additional characteristic positions not listed in Table 2 are put in parentheses.

The diagnostic position 78 (Table 2) is in the middle of the stretch 77–79, which can pair with base 87–89 containing another diagnostic position 88. Bases 77–79 are within the polymerase III promoter region (bases 74–86 in Fig. 1). Correlation between occurrences of complementary bases at positions 78 and 88 suggests the possibility of a weak secondary interaction in this region. Another potential for secondary interaction, already proposed for 7SL RNA (7, 8), exists between complementary bases 69–75 and 89–95. This region includes 3 of the 15 positions distinguishing between the J and S subfamilies and the complementarity is conserved throughout the *Alu* family. The only A·C mispairing has been found in this region in the *Alu*-c sequences. The role of the above hypothetical structures is not clear, although their location suggests involvement in *Alu* transcription. There is also a possibility of a secondary interaction between bases 244 and 245 and bases 271 and 272 that includes bases at positions diagnostic for putative d and e branches of the *Alu* family discussed above.

Table 3 indicates that the average overall similarity between *Alu*-J and the *Alu* consensus sequence in nondiagnostic positions is lower than the average similarity between *Alu*-S and the *Alu* consensus. This indicates that on average *Alu*-J sequences are more diverse than *Alu*-S sequences. By *t* test, one can find that differences between *Alu*-J/consensus and *Alu*-S/consensus similarities are statistically significant ($P << 0.001$). The conclusion holds true even if the general *Alu* consensus is replaced by the *Alu*-J consensus (data not shown). There is also a significant difference ($P < 0.001$) between analogous numbers for a and b subdivisions of the *Alu* sequences. The differences between *Alu*-b and *Alu*-c sequences are marginally significant ($P < 0.05$), and analogous differences between *Alu*-a and *Alu*-c are insignificant.

As pointed out before (4), CpG doublets undergo rapid mutations in *Alu* sequences. This may result from a deamination of methylated cytosine (for a review, see ref. 9). Average CpG content is lowest in the J subfamily (3.84 ± 2.01) as compared to analogous numbers for *Alu*-a (7.75 ± 2.95), *Alu*-b (16.08 ± 5.01), and *Alu*-c (9.54 ± 3.75) branches of the S subfamily. Significance levels for the differences in the CpG content are virtually identical to those for the similarity differences discussed in the preceding paragraph.

DISCUSSION

Given the similarity between *Alu*-J and 7SL RNA sequences in the diagnostic positions, the large intra-subfamily diversity and the low CpG content, we find the J sequences to be good

Table 2. Base preferences in the S subfamily branches

Consensus position	Branches of *Alu*	Frequency of T	C	A	G	(–)
65 (C)	a	20	41	1	1	8
	c	0	0	0	0	11
	b	0	0	0	0	12
66 (T)	a	62	3	3	0	3
	c	4	0	0	0	7
	b	4	0	0	0	7
78 (T)	a	67	0	4	0	0
	c	1	0	9	1	0
	b	0	0	12	0	0
88 (G)	a	2	1	2	65	1
	c	9	0	1	1	0
	b	11	0	1	0	0
95 (C)	a	2	68	1	0	0
	c	2	9	0	0	0
	b	12	0	0	0	0
100 (T)	a	66	1	2	1	1
	c	7	4	0	0	0
	b	1	10	1	0	1
153 (C)	a	10	35	1	24	1
	c	5	1	0	5	0
	b	0	1	0	11	0
163 (A)	a	1	0	53	17	0
	c	1	1	1	8	0
	b	0	0	1	11	0
197 (C)	a	15	50	2	4	0
	c	3	6	1	1	0
	b	0	0	0	12	0
200 (T)	a	65	3	2	1	0
	c	10	0	1	0	0
	b	1	0	4	7	0
219 (G)	a	0	1	2	64	0
	c	0	1	0	10	0
	b	0	11	0	0	0

(–), Alignment gaps.

candidates for the early *Alu* elements derived from the 7SL RNA (1). The base differences in the diagnostic positions and the linker regions may be important for understanding how this transformation occurred and are good targets for experimental analysis. On the other hand, the least diverse *Alu*-b sequences can be viewed as a relatively young branch of the *Alu* family. There are three published examples of *Alu* sequences that are believed to be inserted relatively recently on the evolutionary time scale: in the α-satellite DNA of African green monkey (10), in the gorilla β-globin gene cluster (11), and at the Mlvi-2 locus of human cell lymphoma (12). All these *Alu* sequences belong to the b branch defined above.

While this paper was in review, other authors (13) reported on a subdivision of the *Alu* family into three different subfamilies corresponding to our J subfamily and two branches (a and b) of the S subfamily. These two branches, as well as the branch c, are virtually equally different from the J subfamily of *Alu* sequences and similar to each other in the diagnostic positions from Table 1. Therefore, we consider them as members of the S subfamily. The authors draw their conclusions from analysis of pairwise difference distribution among *Alu* sequences involving both the diagnostic differences discussed in this paper and a mutational noise. Our analysis is based on multiple sequence comparisons, which permits more rigorous distinction between diagnostic and background differences. With this level of resolution we are able to classify each *Alu* sequence individually. This, and the analysis of the CpG content discussed in the accompanying paper (14), opens a way to date the invasion of individual genes by different types of *Alu* sequences and of genetic rearrangements associated with this process.

Table 3. Average overall similarities with the *Alu* consensus sequence

	Gaps as mismatches		Gaps excluded		
Alu type	Mean	SD	Mean	SD	Total
All	83.88	5.63	86.39	4.38	125
J	79.20	4.35	82.83	2.27	31
S	86.59	3.39	88.75	1.98	94
a	86.52	3.26	88.59	1.79	71
c + b	89.20	4.14	91.15	3.08	23
c	87.04	4.25	89.45	2.87	11
b	91.22	2.92	92.71	2.45	12

Sequence alignments have been made by using the computer algorithm (2). The diagnostic positions have been excluded from similarity calculations.

We thank Donald Faulkner for professional computer assistance and Roy Britten, Douglas Brutlag, Terry Friedemann, David Kristofferson, and Randall Smith for critical and useful comments on the manuscript.

1. Ullu, E. & Tschudi, C. (1984) *Nature (London)* **312**, 171–172.
2. Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. & Matsubara, K. (1987) *Gene* **53**, 1–10.
3. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **145**, 195–197.
4. Bains, W. (1986) *J. Mol. Evol.* **23**, 189–199.
5. Faulkner, D. V. & Jurka, J. (1988) *Trends Biochem. Sci.*, in press.
6. Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. & Deininger, P. (1987) *Mol. Biol. Evol.* **4**, 19–29.
7. Gundelfinger, E. D., Di Carlo, M., Zopf, D. & Melli, M. (1984) *EMBO J.* **3**, 2325–2332.
8. Zwieb, K. (1985) *Nucleic Acids Res.* **13**, 6105–6124.
9. Bird, A. P. (1987) *Trends Genet.* **3**, 342–347.
10. Grimaldi, G. & Singer, M. F. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1497–1500.
11. Trabuchet, G., Chebloune, Y., Savatier, P., Laucher, J., Faure, C., Verdier, G. & Nigon, V. M. (1987) *J. Mol. Evol.* **25**, 288–291.
12. Economou-Pachnis, A. & Tsichlis, P. N. (1985) *Nucleic Acids Res.* **13**, 8379–8387.
13. Willard, C., Nguyen, H. T. & Schmid, C. W. (1987) *J. Mol. Evol.* **26**, 180–186.
14. Britten, R. J., Baron, W. F., Stout, D. & Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4770–4774.

Multiple aligned sequence editor (MASE)

Donald V. Faulkner and Jerzy Jurka

Cognitive capacities of the human brain can not, so far, be matched by computers. Even well optimized computer programs have limited flexibility in addressing the variety of problems associated with sequence analysis. Hence, we were motivated to design a Multiple Aligned Sequence Editor (MASE) which combines manual sequence manipulations with standard computer analysis. An earlier article in *TIBS* described the adaptation of standard word processor software for a similar purpose[1].

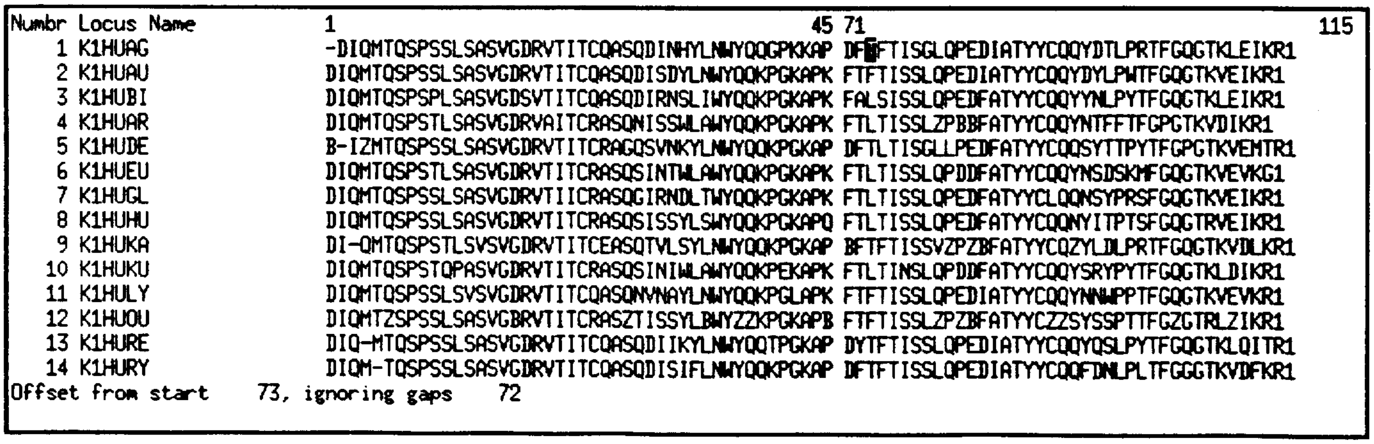
MASE can be used for editing any set of sequences. The total number and size of sequences that can be edited simultaneously depends on the computer (typically, the total number of characters should not exceed 1 × 10^6). Sequences can be displayed on the screen in two windows, each of which can be moved either independently or concurrently. An example of the sequence display is shown in Fig. 1. The window sizes depend upon the type of terminal used. A standard VT100 terminal can display 21 sequences per screen, with 25 characters in each window. Any terminal size and type which supports full screen editing can be used by writing a UNIX termcap entry. MASE can be run on computers with the Berkeley UNIX operating system.

Intrinsic functions

MASE provides over 80 intrinsic functions, each of which can be selected with few key strokes from a menu by calling the function COMMANDMODE(:). Individual functions can also be fixed to different keys on the keyboard using the BIND function, and this assignment can be stored in a separate file and loaded whenever the sequence editor is used. A short definition of each intrinsic function can be called up using question-mark key. Full on-screen HELP and a tutorial are also available. The intrinsic MASE functions can be divided into the following basic groups: (1) moving cursors and searching for patterns; (2) modifying the sequence data; (3) changing the sequence display without affecting input/output files; (4) window manipulations; (5) sequence analysis; (6) modifying MASE behaviour; (7) miscellaneous, and (8) generation of formatted output. In this short article we will outline only some of the capabilities of these functions.

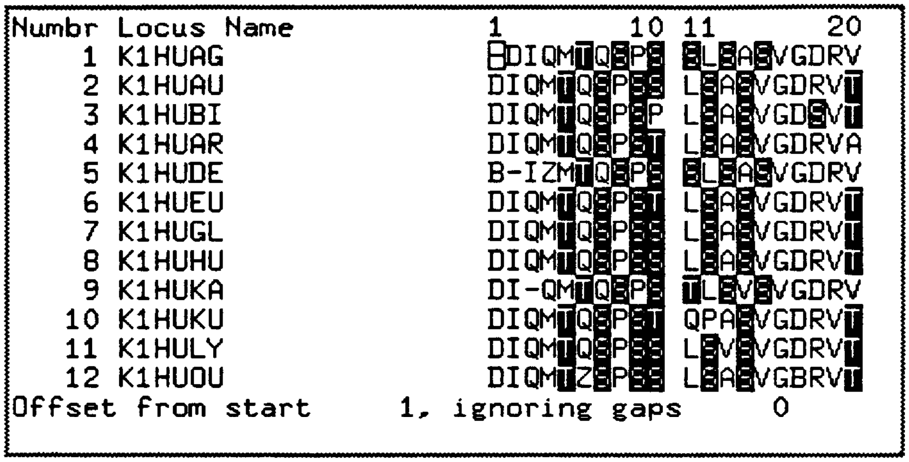
Primary modifications of the sequence data involve insertions/deletions of alignment gaps in individual sequences or in the whole set. This permits the alignment for maximum similarity. One can begin with totally unaligned sequences or use an output from any sequence alignment program as a starting point. However, the format of the pre-aligned sequences must be as described below under the 'Sequence data files' and in the MASE manual. The on-screen alignment can be facilitated by changing the sequence display, for example by highlighting conserved sequence patterns, hydrophobic/hydrophilic residues, etc. (see Fig. 2). One can also easily emphasize differences between aligned sequences. The sequences can be rearranged arbitrarily to facilitate direct by-eye comparisons. Furthermore, the window display permits any two columns of characters be placed next to each other (e.g. columns 45 and 71 in Fig. 1).

Sequence analysis involves on-screen computation of consensus sequence, identity matrix, column com-



```
Numbr Locus Name          1                                          45 71                                          115
    1 K1HUAG              -DIQMTQSPSSLSASVGDRVTITCQASQDINHYLNWYQQGPKKAP DFTFTISGLQPEDIATYYCQQYDTLPRTFGQGTKLEIKR1
    2 K1HUAU              DIQMTQSPSSLSASVGDRVTITCQASQDISDYLNWYQQKPGKAPK FTFTISSLQPEDIATYYCQQYDYLPWTFGQGTKVEIKR1
    3 K1HUBI              DIQMTQSPSPLSASVGDSVTITCQASQDIRNSLIWYQQKPGKAPK FALSISSLQPEDFATYYCQQYYNLPYTFGQGTKLEIKR1
    4 K1HUAR              DIQMTQSPSTLSASVGDRVAITCRASQNISSWLAWYQQKPGKAPK FTLTISSLZPBBFATYYCQQYNTFFTFGPGTKVDIKR1
    5 K1HUDE              B-IZMTQSPSSLSASVGDRVTITCRAGQSVNKYLNWYQQKPGKAP DFTLTISGLLPEDFATYYCQQSYTTPYTFGPGTKVEMTR1
    6 K1HUEU              DIQMTQSPSTLSASVGDRVTITCRASQSINTWLAWYQQKPGKAPK FTLTISSLQPDDFATYYCQQYNSDSKMFGQGTKVEVKG1
    7 K1HUGL              DIQMTQSPSSLSASVGDRVTIICRASQGIRNDLTWYQQKPGKAPK FTLTISSLQPEDFATYYCLQQNSYPRSFGQGTKVEIKR1
    8 K1HUHU              DIQMTQSPSSLSASVGDRVTITCRASQSISSYLSWYQQKPGKAPQ FTLTISSLQPEDFATYYCQQNYITPTSFGQGTRVEIKR1
    9 K1HUKA              DI-QMTQSPSTLSVSVGDRVTITCEASQTVLSYLNWYQQKPGKAP BFTFTISSVZPZBFATYYCQZYLDLPRTFGQGTKVDLKR1
   10 K1HUKU              DIQMTQSPSTQPASVGDRVTITCRASQSINIWLAWYQQKPEKAPK FTLTINSLQPDDFATYYCQQYSRYPYTFGQGTKLDIKR1
   11 K1HULY              DIQMTQSPSSLSVSVGDRVTITCQASQNVNAYLNWYQQKPGLAPK FTFTISSLQPEDIATYYCQQYNNWPPTFGQGTKVEVKR1
   12 K1HUOU              DIQMTZSPSSLSASVGBRVTITCRASZTISSYLBWYZZKPGKAPB FTFTISSLZPZBFATYYCZZSYSSPTTFGZGTRLZIKR1
   13 K1HURE              DIQ-MTQSPSSLSASVGDRVTITCQASQDIIKYLNWYQQTPGKAP DYTFTISSLQPEDIATYYCQQYQSLPYTFGQGTKLQITR1
   14 K1HURY              DIQM-TQSPSSLSASVGDRVTITCQASQDISIFLNWYQQKPGKAP DFTFTISSLQPEDIATYYCQQFDNLPLTFGGGTKVDFKR1
Offset from start    73, ignoring gaps    72
```

D. V. Faulkner is at the Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA, and J. Jurka is at Bionet, 700 East El Camino Real, Mountain View, CA 94040, USA.



```
Numbr Locus Name            1        10 11       20
    1 K1HUAG                BDIQMTQSPS SLSASVGDRV
    2 K1HUAU                DIQMTQSPSS LSASVGDRVT
    3 K1HUBI                DIQMTQSPSP LSASVGDSVT
    4 K1HUAR                DIQMTQSPST LSASVGDRVA
    5 K1HUDE                B-IZMTQSPS SLSASVGDRV
    6 K1HUEU                DIQMTQSPST LSASVGDRVT
    7 K1HUGL                DIQMTQSPSS LSASVGDRVT
    8 K1HUHU                DIQMTQSPSS LSASVGDRVT
    9 K1HUKA                DI-QMTQSPS TLSVSVGDRV
   10 K1HUKU                DIQMTQSPST QPASVGDRVT
   11 K1HULY                DIQMTQSPSS LSVSVGDRVT
   12 K1HUOU                DIQMTZSPSS LSASVGBRVT
Offset from start     1, ignoring gaps      0
```

Fig. 2. A sequence display from MASE with T and S highlighted.

position and search for compensatory mutations using COLUMN-CORRELATION function. COLUMN-COMPOSITION permits analysis of base/amino acid frequencies at homologous sequence positions. COLUMN-CORRELATION lists the total number of simultaneous and non-simultaneous base differences between the aligned sequences in any two sequence positions. Each sequence can be compared either to a reference sequence (order independent), or to its nearest neighbor (order dependent).The probability of simultaneous and non-simultaneous base variations in the whole set is evaluated using the binominal formula. The COLUMN-CORRELATION function has recently been used for secondary structure prediction and classification of human Alu sequences[2].

Sequence data files

Each input file contains one or more sequences. A sequence can contain an indefinite number of comment lines, each having a semicolon in the first column. The first line without semicolon contains the locus name which can not exceed 20 characters, and this is followed by the sequence information lines each not exceeding 95 characters in length. After insertion of alignment gaps the modified sequence set is stored in the same input format. Aligned output file can be created by a separate OUTPUT-ALIGNED function. Line length, number of lines per page, positions of vertical and horizontal gaps, all depend upon the chosen output format.

Program availability

By the end of this year MASE will be made available on-line for the Bionet community. Other non-profit users can obtain the source code for editor written in C and MASE manuals from: Susan Russo, MBCRR, Dana-Farber Cancer Institute, 44 Binney Street, Boston MA 02115, USA. Tel. (617) 732-3746.

Acknowledgements

This software was developed under auspices of Molecular Biology Computer Research Resource which is supported by NIH grant RR02275-03. A portion of the support came froṁ the NIH grant U41-RR-01685-05 for the Bionet resource. We thank Dr Temple Smith for helpful suggestions and strong encouragement for this project.

References

1 Boswell, R. B. (1987) *Trends Biochem. Sci.* 12, 279–280

2 Jurka, J. and Smith, T. F. (1988) *Proc. Natl Acad. Sci. USA* (in press)

Volume 16 Number 2 1988 Nucleic Acids Research

Small cytoplasmic *Ro* RNA pseudogene and an *Alu* repeat in the human α-1 globin gene

Jerzy Jurka, Temple F.Smith[1] and Damian Labuda[2]

Bionet, 700 East El Camino Real, Mountain View, CA 94040, [1]Dana-Farber Cancer Institute, Harvard University, 44 Binney Street, Boston, MA 02115, USA and [2]Medical Genetica, Hopital Sainte-Justine, 3175 Cote Sainte-Catherine, Montreal, Quebec H3T 1C5, Canada

Submitted November 6, 1987

The 5'-end of the previously studied *Alu* repeat from the α1-globin gene (1) is flanked by a sequence 80% similar to one of the full length human small cytoplasmic *Ro* RNAs (Fig. 1a), denoted as HY3 (2). This is the first known example of a pseudogene for the *Ro* scRNA. Only a few such pseudogenes are expected to exist in the human genome (2). The pseudogene location next to the *Alu* sequence may suggest physical interactions between HY3-like RNA and the *Alu* RNA prior to the reverse transcription. The 3'-flanking region of the previously studied full size *Alu* repeat is another unreported *Alu* sequence truncated at the *Eco* RI restriction site (Fig. 1b). *In vitro* transcription of the region analysed (1) gave four RNA fragments. One of them, 86 nt long, is synthesized from the short class III transcriptional unit located on the 5'-side of the *Alu* repeat (3). This location coincides with the location of the HY3-like DNA sequence.

```
promoter?         GTGG-CNNAGTGG
HY3             GGCTGGTCCGAGTGCAGTGGTGTTTACAACTAATTGATCACAACCAGTTA    50
                *******| ********|* * *******|********** *********

3'-α1           GGCTGGTTGGAGTGCAGCGCTTTTTACAATTAATTGATCAGAACCAGTTA    52
                                                                          (a)
HY3            CAGATTTCTTTGTTCCTTCTCCACTCCCACTGCTTCACTTGACT-AGCCTTT  101
               |*|**** *|   ****************||********* ****** *****

3'-α1          TAAATTTATCÀTTTCCTTCTCCACTCCTGCTGCTTCAGTTGACTAAGCCTAA  104

promoter          GTGGCNNAGTGG
Alu            GGCCGGGCGCGGTGG-CTCACGCCTGTAATCCCAGCACTTTGGGAGGCCG     49
               **||****|*|**** ******************************|***|

3'-α1          GGTTGGGCACAGTGGCCTCACGCCTGTAATCCCAGCACTTTGGGAAGCCA    471
                                                                          (b)
promoter                                GGGTTCGANNCC
Alu            AGGCGGGCGGATCACCTGAGGTCAGGAGTTC                        80
               ***|****|******   |*********|***

3'-α1          AGGTGGGCAGATCAC--AAGGTCAGGAATTC                       500
```

Fig. 1. (a) Sequence alignment between HY3 (2) and the corresponding 3'-α1-globin region (1). Putative polymerse III promoter is indicated. (b) Genomic *Alu* consensus (4), aligned to the 3'-α1-globin sequence at positions 423-500. Promoter boxes (5) are indicated. Sequences and numbering of the 3'-α1-globin region are identical to those in (1). Exact matches (*), purine-purine/pyrimidine-pyrimidine replacements (|), and gaps (-) are indicated in both alignments.

REFERENCES

(1) Shen, C.-K.J. and Maniatis, T. (1982) J. Mol. Appl. Gen. 1, 343-360. (2) Wolin, S.L. and Steitz, J.A. (1983) Cell 32, 735-744. (3) Hess, J., et al. (1985) J. Mol. Biol. 184, 7-21. (4) Schmid, C.W. and Shen, C.-K.J. (1985) in Molecular Evolutionary Genetics (McIntyre, R.J. ed.), pp. 323-358. Plenum Publishing. New York. (5) Fowlkes, D.M. and Shenk, T. (1980) Cell 22, 405-413.

EVOLUTION OF HUMAN ALU REPEATS: IMPLICATIONS FOR GENOME STUDIES

J. Jurka[1] and R.J. Britten,[2] [1]Bionet, Mountain View, California; [2]California Institute of Technology, Pasadena, California.

The human Alu family of repeated sequences contains at least five different subfamilies referred to as Alu-b,c,d,e and j which are arranged according to their position on the evolutionary time scale from the most recent, Alu-b, to the oldest, Alu-j, subfamily. This is identified by computer analysis of the correlated, subfamily -specific nucleotide occurrences in a number of the diagnostic Alu sequence positions. A specialised sequence editor has been developed to pursue these studies.

Members of each subfamily show different nucleotide preferences in the diagnostic Alu sequence positions which can easily be identified Another characteristic feature of different subfamilies is a systematic difference in the CpG content from an average 16.08 per sequence in Alu-b to an average 3.84 per sequence in the Alu-j subfamily. It is proposed that subsequent generations of Alu subfamilies have been transcribed from different CpG-rich source genes. The genes presumably have replaced each other during the evolutionary history of primates. Once integrated into the genome, a copy of the source gene is no longer under selective pressure to maintain the original CpG content of the source gene. Therefore, one can observe a time-dependent elimination of CpG from Alu sequences.

The classification of Alu repeats, based on the diagnostic base differences and the CpG "decay", provide a method to date both the invasion of individual genes by Alu elements and the genetic rearrangements associated with this process.

Cold Spr. Harb. Symp., May 1988

STUDIES ON RAT LIVER CYTOCHROMES P450 USING COMPARATIVE SEQUENCE ANALYSIS

Elzbieta Holsztynska[1], David J. Waxman[1] and Jerzy Jurka[2]
(1) Department of Biological Chemistry and Molecular Pharmacology, Dana-Farber Cancer Institute, Harvard Medical School; (2) Bionet, National Computer Resource for Molecular Biology.

Currently, about 70 full-length sequences for cytochrome P-450 from 9 eukaryotic species and one prokaryote are available. The crystallographic model of bacterial cytochrome has been reported (Poulos et al., 1987, J. Mol. Biol. 195, 687) and has been used as a reference to evaluate our structural predictions using a variety of theoretical and computer methods. Postulated regions of potential structural-functional importance in P450 PB4(IIB1) include: (1) combined membrane insert halt-transfer signal residues; (2) sites of interaction with cytochrome b5 and cAMP-dependent kinase; (3) internal halt-transfer sequence; (4) large hypervariable region; (5) predicted dioxygen binding site; (6) NADPH-P450 reductase interaction site; (7) conserved region and (8) axial cysteine heme-binding region. We used Multiple Aligned Sequence Editor (Faulkner and Jurka, 1988, TIBS, in press), to elaborate a synthetic model correlating sequence variations in homologous positions of the aligned set with the functional map as well as with predicted and/ or reported structural features. Results of this study will be presented.

(Protein Soc. Meeting, San Diego, August 1988)

Locating Amino Acid Patterns in Proteins by Composition

Sunil Maulik
BIONET c/o IntelliGenetics, Inc., 700 E. El Camino Real, Mountain View, CA., 94040.

It is known that the amino acid composition of a protein plays a major role in determining its folded state [1]. A fundamental pattern-matching problem in protein analysis is that of finding sequences (or sub-sequences) given certain *compositional* criteria only. For instance, one may wish to find all regions *of unspecified length* >20% proline and >30% glycine residues in a (say) 500 amino acid peptide sequence. A related problem may be to find all hydrophobic regions (>50%) in a protein. An algorithm has been developed that will scan a sequence and find sub-sequences (of any length) satisfying given compositional criteria. The implementation of this algorithm, termed RICH, is near completion and will be available soon on BIONET. Uses of the RICH program might include finding hinge structures in immunoglobulin sequences (known to be rich in proline and cysteine residues [2]); or verifying if a protein satisfies the PEST hypothesis [3] i.e. if its half-life is related to the compositional content of P (proline) E (glutamic acid) S (serine) and T (threonine) residues.

References:

1. Sheridan RP, et al.,(1985) Biopolymers 24: 1995-2003
2. Huber R and Bennett WS (1987) Nature 326: 334-335
3. Rogers S Wells R and Rechsteiner M (1986) Science 234: 364-368

II. BIONET Training Publications

A copy of one BIONET informational publication is included in this section.

Protein Seq Data Anal (1988) 2: ■■-■■

Protein Sequences & Data Analysis
© Springer-Verlag 1988

Protein databases and software on BIONET

Sunil Maulik

BIONET c/o IntelliGenetics Incorporated, 700 East El Camino Real, Mountain, View, CA, 94040 USA

Abstract. BIONET provides databases, software, and networking/communications tools to over 2500 molecular biologists worldwide. Software for the analysis of nucleic acid and protein sequence data is provided by both IntelliGenetics, and academic contributors. BIONET is currently implementing dedicated high speed servers for searching protein databases, as well as providing more flexible tools for protein structure recognition and prediction. In this review, protein databases and analysis software available on the BIONET resource are described, and progress in providing new tools for structure prediction, comparative sequence analysis, and pattern recognition using Artificial Intelligence (AI) techniques are summarized.

What is BIONET?

BIONET is a national computer network for molecular biologists and biochemists. It is a non-profit resource funded by a co-operative agreement between the NIH, Division of Research Resources and IntelliGenetics, of Mountain View, California (No. 5 U41 RR01865-06). BIONET provides access to biological databases, software for analyzing the data, and communication and networking tools for the distribution of data, software, and computing resources. BIONET maintains databases and software for both nucleic acid and protein analysis. This review covers the protein analysis component of BIONET only.

An annual subscription fee of $400 allows academic or non-profit users unlimited access to the BIONET computer (a DEC-2065 time-shared mainframe, with additional network access to a Sun 3/280 database server and a microVAX) including payment of telecommunication costs. Access limited only to the communications facilities is also available. International users are waived the subscription fee, but are required to pay their own telecommunication costs. Over 750 laboratories are currently subscribed to BIONET from the U.S., Europe, and Japan, corresponding to over 2500 researchers.

BIONET is directly connected to Telenet and CompuServe which provide 24-h access to the BIONET computer via a local telephone call from most cities in the United States. These networks allow scientists in Europe and Japan to access BIONET using international carriers such as Euronet, Datax-P, Transpac, and Venus. BIONET is also directly connected to the ARPAnet which provides mail and file-transfer capabilities to any host computer on the ARPAnet and to a large number of Internet hosts on Bitnet, Usenet, CSnet, etc. This extremely high level of connectivity has facilitated both communication and collaboration between scientists worldwide. As an example, Dr. M.M. Teeter of Boston College (Teeter@ bcchem.Bitnet) maintains a database of the electronic mail addresses of protein crystallographers throughout the U.S. and Europe. This database is accessible on BIONET for use with the MM mail software.

BIONET is using its extensive networking and communications facilities to maintain a number of electronic bulletin-boards (BBoards) dealing with different aspects of molecular biology. These BBoards are distributed throughout the world. Messages are exchanged with the SEQNET BBoard in Europe (SEQNET is BBoard service run by Drs. Michael Ashburner and Martin Bishop of Cambridge University), and are received as far away as Israel, Korea, Taiwan, Australia, and even the U.S. research station in Antarctica! Recently, BIONET helped initiate and participated in the formation of the worldwide BIOSCI BBoard network. Of particular interest to scientists working with proteins are the PROTEIN-ANALYSIS and METHODS-AND-REAGENTS BBoards. The former is moderated by Amos Bairoch of the University of Geneva (Bairoch@ cgecmu51.Bitnet) and deals with such topics as protein chemistry, sequence analysis, and structure prediction. The latter is an all-purpose BBoard for nucleic acids and proteins, and contains requests and data on topics such as codon usage tables, peptide synthesizers, cDNA libraries, antibodies, and protein engineering/site-directed mutagenesis.

Protein databases and software

BIONET provides the Protein Identification Resource (PIR) protein sequence database [1], the SWISS-PROT protein sequence database distributed by the European Molecular Biology Laboratory (EMBL)[2], and the KeyBank™ database of protein (and nucleic acid) sequence patterns (provided by IntelliGenetics). KeyBank™ contains protein and DNA consensus sequences (motifs) described using the QUEST programs pattern language [3]. Thus, at a single resource information at several levels of biological function are immediately accessible to the researcher. IntelliGenetics provides a suite of programs that readily access information in these databases. The IFIND program will rapidly search a query sequence against any sequence databank utilizing

Ms. No. 035 Author Maulik Ms. 1 – 15 Pages 1 – 4

Springer-Verlag, Heidelberg / H. Stürtz AG, Würzburg

Provisorische Seitenzahlen / Provisional page numbers I. Korr. P Date 23.3.88

the algorithm of Wilbur and Lipman [4], report similarity scores, and display optimized alignments between the query and similar database sequences. The QUEST program can also search any sequence databank with a given sequence pattern and locate exact matches of that pattern to sequences or subsequences in the database. QUEST may also be used in conjunction with the predefined patterns in Key-Bank'm to locate particular subsequences (for instance, signal sequences) within a given sequence of interest.

Sequences located by either IFIND or QUEST may be simultaneously aligned by GENALIGN, a multiple sequence comparison program utilizing the regions method of Martinez [5]. GENALIGN will align up to 49 protein or nucleic acid sequences simultaneously. The program allows the user to choose between several different amino acid "alphabets" when comparing protein sequences, including the Jimenez-Montano and Zamora-Cortina alphabet of evolutionary similarity [6], the Miyata alphabet of physico-chemical similarity [7], a hydrophobic-hydrophilic alphabet, or a hydrophobic-neutral-hydrophilic alphabet. In addition, users have a flexibility to create and use their own amino acid alphabets.

IntelliGenetics' PEP program allows the user to perform various analyses on peptide sequences. These include secondary structure predictions by the Chou-Fasman algorithm [8], and hydropathicity calcuations using the Hopp-Woods [9] or Kyte-Doolittle [10] procedures. In addition, PEP can simulate chemical cleavage by a variety of proteases and chemical treatments (a database of eight common proteases and five forms of chemical cleavage exist within the program[1], and users can add or create their own database), determine amino acid composition, molecular weight, and pI. The program also performs rapid and rigorous similarity comparisons between peptide sequences. (The former using a modified [11] version of the Korn-Queen-Wegman algorithm [12], and the latter using the Needleman-Wunsch algorithm [13] as modified by Smith-Waterman [14].) PEP determines reverse translations using codon preference tables followed by restriction site mapping and splicing functions. Finally, PEP also contains a generalized "window" algorithm, that allows users to define any characteristic pattern in a sequence that can be determined using a moving weighted average (either arithmetic or geometric) over the sequence. For example, the window function in PEP may be customized to predict the existence of membrane-associated alpha-helices using the algorithm of Rao and Argos [15].

In contrast to the other programs in the IntelliGenetics suite, the SIZER program does not deal directly with sequence data, but instead may be used to calibrate gel electrophoresis bands against given standards, SIZER can use either the Duggleby [16], Southern [17], or spline [18] methods of curve fitting. SIZER may be used with either protein or nucleic acid gel fragment data.

In addition to the IntelliGenetics suite of programs for nucleic acid and protein sequence analysis, BIONET provides selected software from academic researchers. Contributed programs that deal with various aspects of protein analysis currently on BIONET include:

(i) FASTP, which utilizes the algorithm of Lipman and Pearson [19] to implement rapid and sensitive similarity searches of the PIR or SWISS-PROT protein databanks

(ii) The PROT3 [20], ALP3 [21] and XALIGN [22] programs for multiple sequence alignment of protein sequences

(iii) The XPROF program for the prediction of protein hydropathicity using Rose's algorithm [23]

(iv) The IDEAS suite of structural programs [24] including DELPHI (secondary structure prediction by Robson's method [25], HPLOT (distribution of hydrophobic and charged residues using the Nozaki-Tanford [26] or Eisenberg et al. [27] methods), HCOMP (comparison of hydrophobicity profiles), and ALOM (prediction of membrane-spanning regions by discriminant analysis [28]).

(v) The DSSP program [29] for the determination of protein secondary structure from the Brookhaven Protein databank atomic coordinates.

BIONET also creates software which promotes efficient use of the resource. Recent software additions include the BIFIND and BFASTP database interactive command-file generators. BIFIND and BFASTP serve to insulate naive users from the intricacies of performing database similarity searches. They prompt the user for sufficient information to perform the database search, display menus of the different databases, and use intelligent defaults for all other parameters. They produce as their output command (batch) files, which, when submitted, instruct the appropriate database searching program (IFIND in the case of BIFIND, FASTP in the case of BFASTP) to run the searches as batch (remote) jobs.

New directions

An extremely rapid version of FASTP implemented to run on Sun workstations has been provided to BIONET by Dr. Warren Gish of U.C. Berkeley. Named FASTP-mail, it is accessible by sending an electronic mail message containing the query sequence to an IntelliGenetics Sun database-server. The program automatically reads the message and then proceeds to search the entire PIR database in as little as 30 seconds. It remails the search output (top 20 scoring database sequences as well as optimized alignments) back to the user. The implementation allows the user to proceed with other tasks on BIONET, or even log-off, while the search is taking place.

BIONET is implementing a FASTA-mail program (that searches both the GenBank/EMBL nucleotide libraries and the PIR/SWISS-PROT peptide libraries). Thanks to a generous equipment donation from Sun Microsystems, BIONET will be moving to a network of Sun workstations and file-servers during 1988. The new network will result in a greater than tenfold increase in the amount of total compute power available on BIONET. The implementation of dedicated database-servers on this network will become increasingly important as both the resource and the databases continue to expand.

BIONET has initiated a collaboration with RIACS (Research Institute for Advanced Computer Science) to use the Connection Machine II [30] for database searches. The Connection Machine II is an example of a massively parallel architecture. BIONET intends to implement software

[1] The proteases are: trypsin, chymotrypsin, pepsin, thermolysin, clostripain, Staphylococcal protease, Myxobacter protease, and proendopeptidase. The chemical treatments are: cyanogen, bromide, hydroxamine, iodosobenzoate, pH 2.5, and iodoacetamide

for rapid database similarity searches and large-scale sequence alignments within this parallel environment.

In the last fifteen years protein structure prediction has become an increasingly important research topic. There is a growing demand from the BIONET community for more flexible and accurate tools for protein structure prediction and analysis. BIONET intends to build upon its expertise in comparative sequence analysis [31] and to collaborate with outside experts on protein structure recognition, analysis, and prediction to research and develop new tools in this field. Protein structure prediction by knowledge-based analysis requires as its starting point a database of structural knowledge [32] obtained from the analysis of known protein structures. Rules inferred from such a database are then combined with either pattern-matching [33] or multiple sequence alignment algorithms [34] to predict secondary structural elements, and, optimally, a single tertiary structure. If the protein sequence shares a substantial primary sequence similarity (>50%) to a sequence of known structure, multiple sequence alignment, followed by molecular modeling, generally allows a plausible structure to be built [35, 36]. Even in the absence of known structure, predicted secondary structural elements, hydropathicity, chain flexibility, and evolutionary information combined in a multiple sequence alignment can provide increased accuracy in predicting the fold of a protein [36–38]. As obtaining optimal multiple sequence alignments algorithmically remains an unsolved problem, many researchers prefer the interactive alignment and analysis of sequences. A prominent example of software capable of such a task is the Multiple Aligned Sequence Editor (MASE) developed by Faulkner and Jurka [39], which combines interactive sequence manipulations together with standard sequence analysis functionality. MASE runs on computers under the UNIX operating system. It will be available on-line later this year when BIONET moves to the SUN network.

A complementary approach to protein structure prediction by comparative sequence analysis is one of pattern-matching. The QUEST program described earlier already provides one means of finding sophisticated patterns in sequences. However, to recognize structural features of proteins, more complex pattern searching tools are needed. A pattern language for protein structure (termed PLANS) has already been implemented [40]. An artificial-intelligence based program called MATCH has been written by R. Abarbanel. MATCH can find elements of secondary structure such as turns with better than 90% accuracy [33, 41]. Tertiary structure predictions using both pattern-matching and combinatorial approaches are now being attempted [42, 43]. BIONET intends to make MATCH and other programs developed for combinatorial prediction of protein tertiary structure available on the resource. Many of these programs result from the research of F.E. Cohen and I.D. Kuntz at UCSF [33, 40, 41] with whom BIONET intends to collaborate on disseminating research software tools. The programs as well as the knowledge-bases upon which they depend will be expanded to improve the accuracy of predictive techniques.

A related pattern-matching problem is that of finding sequences (or sub-sequences) given certain compositional criteria only. For instance, one may wish to find all regions (of unspecified length) >20% proline and >30% glycine residues in a 500 amino acid peptide sequence. A related problem may be to find all hydrophobic regions (>50%) in a protein. An algorithm has been developed that will scan a sequence and find sub-sequences (of any length) satisfying given compositional criteria. The implementation, termed RICH, is in progress. (Details of the algorithm and its implementation will be described elsewhere.) Uses of the RICH program might include finding hinge regions in immunoglobulin sequences (known to be rich in proline and cysteine residues [44]); or verifying if a protein satisfies the PEST hypothesis [45], i.e., if its half-life is related to the compositional content of P (proline) E (glutamic acid) S (serine) and T (threonine) residues.

Summary

With efforts underway to sequence complete genomes, the number of actual and deduced protein sequences will increase rapidly. Sequence databases and software must become integrated with "higher order" databases (e.g., of protein structural characteristics). Suitable software must also be developed to link them together. BIONET already acts as a central repository for sequence databases, sequence and pattern searching software, and electronic media for the acquisition and dissemination of new biological information. Additionally, BIONET is actively pursuing research to create new algorithms to allow the functional characteristics of protein molecules to be deduced from their sequences.

References

1. George DG, Baker WC, Hunt LT (1986) NAR 14:11–16
2. Hamm GH, Cameron GN (1986) NAR 5–10 14.
3. Abarbanel RA (1984) NAT 12:263–280
4. Wilbur WJ, Lipman DJ (1983) Proc Natl Acad Sci USA 80:726–730
5. Sobel F, Martinez HM (1986) NAR 14:363–374
6. Jimenez-Montano M, Zamora-Cortina L (1981) Proceedings, VII International Biophysics Congress, Mexico City
7. Miyata T, Miyazawa S, Yasunaga T (1979) J Mol Evol 12:219–236
8. Chou PY, Fasman GD (1974) Biochemistry 13:211–245
9. Hopp TP, Woods KR (1981) Proc Natl Acad Sci USA (1981) 78:3824–3828
10. Kyte J, Doolittle RF (1982) J Mol Biol 157:105–119
11. Brutlag D, Clayton J, Friedland P, Kedes LH (1982) NAR 10:279–294
12. Korn LJ, Queen CL, Wegman MN (1977) Proc Natl Acad Sci USA 74:4401–4405
13. Needleman SB, Wunsch CD (1970) J Mol Biol 48:443–453
14. Smith TF, Waterman MS (1981) J Mol Biol 147:195–197
15. Rao MJK, Argos P (1986) Biochem Biophys Acta 869:197–214
16. Duggelby R, Kinns H, Rood JI (1981) Anal Biochem 110:49–55
17. Southern EM (1979) Anal Biochem 100:319–323
18. Vandergraft JS (1983) Spline interpolation. In: [redacted] (eds) Introduction to numerical computations, 2nd ed. Academic Press, San Francisco, pp 126–138
19. Lipman D, Pearson W (1985) Science 227:1435–1441
20. Murata M, [redacted] (1985) PNAS 82:3073–3077
21. Gotoh O (1986) J Theor Biol 121:327–337
22. Bacon DJ, Anderson WJ (1986) J Mol Biol 191:153–161
23. Rose GD, [redacted] Science (1985) 229:834–838
24. Kanehisa M (1984) IDEAS. Integrated database and extended analysis system for nucleic acids and proteins. User manual. Laboratory of Mathematical Biology, NIH, Bethesda, MD
25. Garnier J, Osguthorpe DJ, Robson B (1978) J Mol Biol 120:97–120
26. Nozaki Y, Tanford C (1971) J Biol Chem 246:2211–2217

4

27. Eisenberg D, Weiss RM, Terwilliger TC (1984) Proc Natl Acad Sci USA 81:140–144
28. Klein P, Kanehisa M, DeLisi C (1985) Biochem Biophys Acta 815:468–476
29. Kabsch W, Sander C (1983) Biopolymers 22:2577–2637
30. Hill D (1985) The connection machine. MIT Press, Cambridge, MA .15
31. Jurka J, Smith TF (1988) Proc Natl Acad Sci USA 85:4775–4778
32. Blundell TL (1987) Nature 326:347–352
33. Cohen FE, [redacted] [redacted] [redacted] (1983) Biochemistry 22:4894–4904
34. Webster TA, Lathrop RH, Smith TF (1987) Biochemistry 26:6950–6957
35. Lesk AM, Chothia CH (1986) Phil Trans R Soc London Ser A 317:345
36. Brown JH, [redacted] [redacted] [redacted] (1988) Nature 332:845–850
37. Crawford IP, Niermann T, Kirschner K (1987) Proteins 118–129
38. Pearl LH, Taylor WR (1987) Nature 329:351–354
39. Faulkner DV, Jurka J (1988) Trends Biochem Sci [redacted]:[redacted]–[redacted]
40. Abarbanel RA (1984) Ph.D. thesis, University of California, San Francisco
41. Cohen FE, [redacted] [redacted] [redacted] (1986) Biochemistry 25:266–271
42. Cohen FE, [redacted] [redacted] [redacted] (1986) Science 234:349–352
43. Webster TA, Lathrop RH, Smith TF (1987) Biochemistry 26:6950–6957
44. Huber R, Bennett WS (1987) Nature 326:334–335
45. Rogers S, Wells R, Rechsteiner M (1986) Science 234:364–368

Received May 31, 1988 / Accepted June 27, 1988

III. BIONET Newsletters

Copies of the two BIONET Newsletters are included in this section.

BIONET NEWS

vol. 1 no. 1 April 1988

Automatic Data Submission to GenBank, EMBL, and NBRF-PIR

BIONET users can conveniently submit sequence data using the XGENPUB program. XGENPUB helps to annotate and electronically submits sequence data to GenBank, the National Institutes of Health DNA sequence library, EMBL, the nucleotide sequence database from the European Molecular Biology Laboratory, and NBRF-PIR, the National Biomedical Research Foundation's protein sequence database.

Sequence authors are encouraged to submit their sequence data once it has been derived. Direct computer-readable author submissions decrease the time delay with which the data is incorporated into the databases and can improve the completeness and accuracy of the annotation and sequence data.

XGENPUB is located on-line and can be accessed by simply typing "XGENPUB" at the system @ prompt. The program accepts sequence data in files from the author's BIONET directory and prompts the user for the name of the sequence file and the sequence name. XGENPUB initiates an editor for completing a submission form and inserts the sequence data at the end of the form. When the user exits from the session, the completed entry is automatically mailed to GenBank, EMBL, and NBRF-PIR using the ARPANET computer network. A copy of the submission will also be sent to the author's mail file. Sequence data will be present in the databases within four months of submission.

The vast rate of growth that databases are experiencing moves sequence submission responsibilities to the authors. Many journals are now requesting that their contributing authors submit the data. One journal, *Nucleic Acids Research*, requires evidence that data have been submitted before it will consider a manuscript for review. The XGENPUB program provides BIONET users with a convenient method with which to fulfill their sequence submission responsibilities. Further information regarding XGENPUB may be obtained by viewing the on-line help topic HELP XGENPUB. ◘

Efficient Database Searching

In an effort to increase efficient utilization of the resource, BIONET has produced "Command File Generators." These programs create a formatted list of program commands for remote operation of the sequence database searching software on BIONET. BIFIND, BFASTP, and BFASTN serve to insulate users from the intricacies of the IFIND, FASTP and FASTN database similarity searching programs. Each prompts the user for sufficient information to perform the similarity search, displays menus of the different database choices, and uses intelligent defaults for all other parameters.They then produce as their output command (batch) files, which, when submitted, instruct the appropriate database searching program to run the search as a batch or remote job . The user may then safely log-off the BIONET system, knowing that the output of the database search will be stored in a file for subsequent analysis once the batch job has run. On-line help exists for all the programs by typing HELP followed by the program name.

The Command File Generators undergo continued development. As an example, BIFIND will soon be able to save BOTH search and alignment results. With the advent of BIONET's new FASTP-mail database server (that searches the entire PIR protein databank in as little as 30 seconds - 3-5 times faster than any other implementation of FASTP), a new BFASTP is being implemented, that makes use of this new server software. Watch also for a FASTN-mail program capable of searching the GenBank nucleic acid sequence database far more rapidly than previously capable, and a corresponding new BFASTN program. ◘

Revised Introduction to BIONET

In May, a revised version of the *Introducion to BIONET* will be sent to all of the Principal Investigators subscribing to BIONET. The *Introduction to BIONET* is the handbook sent to new subscribers which provides basic information for accessing and using BIONET.

The new version includes two new sections which should benefit users greatly. A "Troubleshooting" chapter helps users diagnose and solve commonly encountered technical problems. A "Program Examples" appendix provides annotated examples of thirteen common biological tasks and how the BIONET programs may be used to solve them. Many more program examples will be available in printed form when the new IntelliGenetics User Manual, described in a separate article, is distributed to BIONET users this spring.

Revisions to the *Introduction to BIONET* include the addition of a step-by-step example of a BIONET user's first login session. It covers how to connect, log into the BIONET account, read login messages, and get help. The chapter describing the BIONET programs now includes a complete list of the contributed molecular biology software that is available on-line or through the BIONET Lending Library of diskettes. The chapter on databases has been expanded to reflect the new organization of the databases as well as recently added databases such as SWISS-PROT. This chapter also describes the new XGENPUB program which allows electronic submission of sequence data and annotations to all of the principal datatbases. Much effort has been given to incorporating user suggestions for changes into the new Intro. We hope that it allows users to spend less time learning, and more time using the BIONET resource. ◘

BIONET News is published by the BIONET Computer Resource staff. Contributors for this issue were Kathryn Berg, Vickie Johncox, David Kristofferson Sunil Maulik and Spencer Yeh. Your comments and suggestions on this publication would be greatly appreciated.

BIONET Communications Reach Worldwide

The BIONET National Computer Resource for Molecular Biology is far more than merely a facility to analyze DNA sequences. In addition to its analysis software and molecular biology databases, the Resource offers access to powerful electronic communications networks with worldwide scope. Use of this facility is completely free to BIONET account holders.

BIONET maintains the most extensive network available of molecular biology electronic bulletin boards (bboards). Nineteen bulletin boards are maintained on the BIONET mainframe computer and copies of posted messages are read around the world. The BIONET bboards provide scientists with the means of contacting a large community of colleagues by simply sending a single electronic mail message. A detailed list of bboards can be seen on BIONET by typing HELP BB-LIST after the @ prompt.

BIONET bulletins are distributed via the ARPANET, BITNET, and USENETcomputer networks and can be received by scientists without BIONET accounts who have electronic mail addresses on any of these networks.There is no charge for this service. Users on any of these networks can also post messages directly to any of these bboards without editorial intervention. One simply uses the address format BBOARDNAME@BIONET-20.ARPA. BIONET users can post messages to the bboards by simply entering the bboardname in response to the To: prompt when sending a message in the MM mail program.

Copies of each message are automatically relayed around the world via our direct network contacts and through our message exchange agreement with the SEQNET bboard service in Cambridge, England. Messages are read by scientists in places as far away as Japan, Korea, Taiwan, Israel, and Australia.

If there are many people working on a local campus computer at your institution who may be interested in participating in the bboards, please have the person in charge of your computer facility contact kristofferson@bionet-20.arpa. We can then arrange an efficient redistribution scheme at your site.

Electronic bulletin boards are just beginning to be discovered by the molecular biology community. Bboards truly have the potential to revolutionize the way that science is done in this field! □

Addressing Electronic Mail Outside of BIONET

Many BIONET users have colleagues with BITNET or EARN (European Academic Research Network) electronic mail addresses and wonder how to send mail to them. The procedure is extremely simple, and there are no charges for sending mail messages. One sends a message using the normal BIONET mail routine; only the address for the message is slightly different. Suppose that a colleague's BITNET address is SMITH@YALEVM. Enter SMITH@YALEVM.BITNET in response to the To: prompt in the MM mail program:

```
@mm
MM> send
To: smith@yalevm.bitnet
```

The procedure is the same for EARN addresses. SMITH@EMBLbecomes SMITH@EMBL.EARN. Actually SMITH@EMBL.BITNET will also work for EARN addresses, since the two networks are essentially the same. EARN is BITNET in continental Europe.

Some users have trouble addressing mail to English addresses. The network used in England is called JANET (Joint Academic NETwork). Unfortunately, JANET addresses are arranged in reverse order of that used in many other countries, including the US. To circumvent this problem, BIONET has implemented the .JANET address extension (properly termed a "pseudo-domain"). To send to a JANET address mad3@uk.ac.cam.bio, add .JANET to the end of the address:

mad3@uk.ac.cam.bio.janet

Note that case does not matter in mail addresses.

Finally, sometimes one encounters troublesome addresses that contain !'s: foop!sop!net@pri.com. When entering these addresses at the To: prompt, enclose the part of the address to the left of the @ in quotes. Enter:

"foop!sop!net"@pri.com

All of the !'s should be enclosed between the two quotation marks for the address to be properly interpreted by the BIONET mail software.

This short article should provide the means to contact users almost anywhere in the networked world provided one has their electronic address. Of course, as with regular mail, one somehow has to find that address by other means. International user directories are still in their infancy, but there are means of determining addresses if the telephone or regular mail can not be used to find this. The BIONET consultants can help with this and any other communications questions. Please do not hesitate to send them an e-mail message to the BIONET address or call them at (415)324-4363. □

User Manual Replaces Short Course

A User Manual, which replaces the current Short Course, will be made available to BIONET Principal Investigators this Spring. Each Principal Investigator will receive one copy of the manual.

The User Manual is organized into the following sequence analysis topics:

- sequence entry and editing
- sequence location
- sequence and map display
- sequencing project management
- sequence translation
- sequence composition
- sequence structure
- restriction analysis
- cloning simulation
- sequence comparision and alignment

Each topic contains annotated program examples illustrating the step-by-step procedures involved in performing specific operations. The User Manual also includes sections on file structure, databases, and program commands. Additional information may be obtained from Kathryn Berg at (415)962-7337. □

Subscription Fees

Each year BIONET account holders are assessed a flat $400.00 fee used to cover telecommunications charges. New accounts are billed when the BIONET account is activated. For renewing subscribers, a reminder notice is sent out, followed by an invoice a few weeks later. It would be appreciated if the invoices for both new accounts as well as renewing subscribers would be paid within a reasonable length of time.

We have instituted a policy of freezing delinquent accounts if the yearly subscription fee is not forthcoming. If your account has been frozen, please call (415)962-7337. □

BIONET NEWS

vol. 1 no. 2 October 1988

BIONET Grant Renewal

The grant for the BIONET National Computer Resource for Molecular Biology is currently up for renewal. We have been advised by the NIH that it is appropriate for us to solicit testimonial letters about the Resource from our user community. While we have collected many comments sent in by users over the last five years, formal testimonial letters are preferred. We request that you take the time as soon as possible to send us your written comments about BIONET (signed hardcopy, please!). Please detail both the positive and negative aspects of the Resource to enable us to assess the service that we provide. Our mailing address is:

Dr. David Kristofferson
BIONET
c/o IntelliGenetics
700 E. El Camino Real
Mountain View, CA 94040

We would also appreciate three copies of any publications from your laboratory since December 1, 1987 that have entailed the use of BIONET. Thank you in advance for your cooperation.

New FASTA-MAIL Program Speeds Similarity Searches

As the first step towards utilizing the new BIONET Sun computers, the FASTA-MAIL program, developed by BIONET's systems programmers (Eliot Lear and Rob Liebschutz), was released August 17. Since then, over 800 GenBank, PIR, and SWISS-PROT searches have been run by BIONET users. The FASTA-MAIL program allows a user to send a mail message containing a nucleic acid or protein query sequence to a BIONET Sun computer. A sequence similarity search is then performed against a nucleic acid or protein databank using the FASTA program developed by William Pearson and David Lipman. The results are sent to the user's mail file on the BIONET DEC-2065 computer. The FASTA program is an improved version of the older FASTP and FASTN programs. By incorporating distinct scoring matrices, both protein and DNA searches are executed by the same program. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. Thus gap-containing sequences score higher on the first pass, and are retained for further consideration and optimized alignment.

To use FASTA-MAIL on BIONET, the query sequence must adhere to standard IntelliGenetics file format conventions. Because FASTA-MAIL is a mail server, two other stipulations also apply: (1) no line in the file can be longer than 80 characters and (2) there can only be one sequence in the file. To start the FASTA-MAIL program, type FASTA-MAIL at the BIONET "@" prompt. The program is quite simple and user-friendly, but we do urge users to read the on-line HELP FASTA-MAIL help topic at the "@" prompt. This documentation explains the FASTA scoring procedure, gives practical advice about submitting jobs, and contains bibliographic references for the program. A second help topic, HELP EX-35, provides a complete step-by-step example of the program from submission of a FASTA-MAIL job through a review of the results. Parameters that the user chooses are which database to search and the value of "KTUP". The KTUP value is analogous to the "WORDSIZE" parameter in the IFIND program. It determines how many consecutive residues must match in the first pass before the region is considered by the program. A lower KTUP value increases the sensitivity of the search, but lengthens the search time exponentially. The turnaround time on searches depends primarily on how many other jobs are in the queue and the KTUP value. Currently, jobs seem to be running in 10 to 20 minutes for protein searches and in 1 to 3 hours for full GenBank searches.

Many users have asked questions regarding the interpretation of search results. Bear in mind that a high similarity score cannot prove biological homology, it can only provide evidence. FASTA-MAIL provides three scores for each sequence displayed, INITN, INIT1, and OPT. INIT1 is the same as the old FASTP or FASTN "INIT" score and is primarily included in FASTA-MAIL for comparison with results from these older programs. The INITN score is the score used on the first pass to rank sequences or discard those below a cutoff score. While the INITN score does allow gaps within matching regions, the gap-penalty does not vary with distance at this point, nor does the program check to see if there is a better alignment of the two sequences yielding a better score. These are included when the final OPT score is calculated by a Needleman-Wunsch/ Smith-Watterman-type of alignment on the region surrounding the highest scoring initial region. The alignment displayed is based on this final alignment procedure. Thus, while the INITN score determines which sequences are kept and the order in which they are displayed, the OPT score gives a more precise gauge of the similarity of sequences. However, since the OPT score is not calculated for each sequence in the database, there is no mean and standard deviation generated for the OPT scores. The INITN score still must be used when one compares the high scoring sequences against the full population.

Two methods are useful for determining statistical significance. One can simply compare the INITN score of the sequence in question with the mean and standard deviation of all INITN scores to see how many standard deviations above the mean the score lies. However, one must keep in mind that in a sufficiently large database search there usually will be some random matches that still score even three or four standard deviations above the mean. To see if a high scoring sequence is truly similar to the target query, Lipman and Pearson suggest randomizing one of the two sequences and then scoring it with the same INITN or OPT scoring procedure. If the original unrandomized score is much higher than randomized runs, this indicates that the high score is related to the sequence order and not just to the sequence composition. William Pearson has developed a program, RDF2, which performs this Monte Carlo simulation. This program will be available on BIONET when users are shifted to the Sun system. Contact Dr. Pearson directly if you wish to obtain RDF2 for PC's or mainframe Unix computers.

The original paper by William Pearson and David Lipman describing the FASTA program, "Improved tools for biological sequence comparison," appeared in the April 1988 issue of PNAS (Vol. 85, pp. 2444-2448). William Pearson may be contacted by e-mail at "wrp@virginia.edu" or at the Department of Biochemistry; Box 440, Jordan Hall; Univ. of Virginia; Charlottesville, VA 22908. BIONET would like to thank William Pearson and David Lipman for allowing BIONET users access to their programs and for providing details regarding the program's scores. We look forward to receiving comments and suggestions about the FASTA-MAIL program.

Multiple Aligned Sequence Editor (MASE)

Analysis of large sequences or large number of sequences can be quite tedious, frustrating and, above all, time-consuming. Most algorithms can only handle sequences of limited size. Outputs from available multiple sequence alignment algorithms often need further refinements to make biological sense. These and related problems with sequence analysis stimulated my joint effort with Don Faulkner to design and develop a general-purpose multiple sequence editor now called MASE. Our work began back in 1986 when I was a research fellow at Harvard. It has been continued as a collaborative project since I joined BIONET as a scientist in the middle of 1987. An introductory article on MASE can be found in the August issue of Trends in Biochemical Sciences, (Faulkner, D.V. & Jurka, J.; vol. 13, pp. 321-322). A limited number of reprints of this article are still available from either co-author.

MASE can be installed on computers running the Berkeley UNIX operating system. Many non-profit research groups with access to a Berkeley Unix system have already acquired the editor free of charge from Harvard. For those non-profit researchers who do not have access to the appropriate hardware, BIONET is authorized to distribute MASE on-line. Currently, BIONET is moving from our DEC-20 to a network of donated SUN computers. After this transition, MASE will be made available on-line in late 1988 or early 1989. From the technical point of view it would be very useful to know how many potential BIONET subscribers need the sequence editor. We therefore ask potential new users to mail an electronic message to jurka@bionet-20.bio.net. Thank you for your help.

Jerzy Jurka

November Training

Are you a new user of BIONET? Or are you an experienced user wanting to become more proficient with your use of BIONET? Plan to participate in an upcoming BIONET training session.

The next class is scheduled for November 17-18, 1988. Topics covered will include sequence data entry and editing, sequencing gel management, nucleic acid and peptide sequence analysis, database structure and sequence retrieval, collecting sequences and searching for patterns, sequence similarity searches, and electronic mail.

The training session will run from 8:30 am until 5:00 pm and from 6:30 until 9:00 pm on November 17th. The evening session will cover the TOPS20 operating system, communications software, the on-line help facilities, as well as an overview of contributed software. The session on November 18th begins at 8:00 am with a question and answer section. Formal instruction begins at 9:00 am and concludes at 5:00. Lunch is provided each day. Evening meals are not provided, but there are reasonably priced restaurants within walking distance.

BIONET is located in the IntelliGenetics Building, 700 East El Camino Real (at the intersection of Highway 85) in Mountain View, CA. This is within easy driving distance from either the San Francisco or San Jose airports.

The cost of the training is kept low to encourage users to attend. The regular price is $100 for the two day program. To help offset the additional cost of transportation, out-of-state subscribers pay $50.00.

The training is limited to 12 people on a first-come, first-serve basis. To reserve a place, call Kathryn Berg at (415) 962-7337. We accept purchase orders, MasterCard, VISA, or personal check. Payment should be sent to:

BIONET
c/o IntelliGenetics
700 East El Camino Real
Mountain View, CA 94040

We are enthusiastic about the BIONET resource and look forward to facilitating your use of it.

1989 Training Schedule

Listed below are the dates for BIONET training in 1989. Note that the course has been expanded to cover additional topics, such as contributed software, file transfers and telecommunications.

February 6, 7 *
March 15, 16, 17
May 17, 18, 19
July 19, 20, 21
September 20, 21, 22
November 15, 16, 17
*evening session on 2/7

BIONET News is published by the BIONET Computer Resource staff. Contributors to this issue were Kathryn Berg, Karen Davis, Jerzy Jurka, David Kristofferson, and Spencer Yeh. Your comments and suggestions on this publication are greatly appreciated.

IV. BIONET Software Lending Library Catalog

A copy of the catalog is provided in this appendix.

Version 1.5/KB/11-88

PC-Lending Library Software

Please send one disk for each program requested. Include a self-addressed, stamped mailer along with the disks and request form. This will greatly streamline the ordering procedure. Send your request to BIONET, 700 E. El Camino Real suite 300, Mountain View, CA 94040. Substantive questions about the programs, or any difficulties with program functions should be addressed to bionet@BIONET-20.arpa.

I. Communications/Terminal Emulation

Multipurpose public-domain software for logging on to BIONET, VT100 emulation and file transfer. Various formats are available:

- Kermit--for IBM PC compatible computers
- MacIIKermit and Kermit Shareware--for Apple Macintosh, Macintosh II, and Macintosh SE computers (send 2 disks for these programs)

II. Editor

- Micro-emacs--text editor available for BIONET subscribers. Either high or low density versions may be requested.

III. Contributed Software

All programs come on 5 1/4 " diskettes

SEQAIDII
D. J. Roufa and D. D. Rhoads
Division of Biology
Kansas State University
Manhattan, KS 66506

Droufa@BIONET-20.arpa

SEQAIDII is a multifunctional program for DNA sequence analysis.

System requirements: IBM PC compatible computer. Working files require 200 000 bytes of disk storage space.

Files Included:

- INSTALL.DOC--Documentation file
- SEQAIDFD.EXE--Self-extracting arechive of SEQAIDII files
- SEQAIDII.NEW--Release notes for version 3.0
- README.DOC--Installation instructions

PCZUCKER

Michael Zuker
National Research Council
Biological Sciences
Room 3115
100 Sussex Dr.
Ottawa, ON K1A 0R6 CANADA

Zuker@BIONET-20.arpa

PCZucker is a program for global prediction of RNA secondary structure.

System requirements: IBM PC compatible computer

The programs require a minimum of 512 kbytes of RAM, version 2.0 (or later) of DOS, and a floppy containing 360 kbytes, or 1.2 mbytes.

Files included:

- Two versions of PCFOLD
    1. PCFOLD.EXE--packs the arrays, to extend the length of the longest fragment it can fold (425 bases maximum).
    2. PCFOLD2.EXE--does not contain packing, can handle only fragments of 345 bases maximum.
- README.DOC--Documentation and explanation of program use
- MENUDAT,MENUDAT2--Data files needed by the program
- FOLD.ENR--Energy file
- FOLD.BAT--Batch file to run PCFold program
- FOLD2.BAT--Batch file to run PCFold2 program
- PSTV--Default sequence file

The source code is available in:

- SOURCES.1--Source programs for PCFOLD.EXE
- SOURCES.2--Source programs for PCFOLD2.EXE

MOLECULE

John Thompson
Carnegie-Mellon University
Biological Sciences
616 Mellon Institute
4400 Fifth Ave
Pittsburgh, PA 15213

Woolford.Thompson@BIONET-20.arpa

This disk contains compressed files for John Thompson's MOLECULE program for display of secondary structure prediction. The programs read .CT files produced by Zuker's PCFOLD program, and display a 2D graphic representation of the structure. Three versions of MOLECULE are available, but only one need be accessed, depending on the monitor-type being used.

System requirements: IBM PC compatible computer

Files Included:

- CMOLECULE.ARC--IBM colorgraphics monitor
- EMOLECULE.ARC--EGA monitor (Enhanced Graphics Adaptor)
- HMOLECULE.ARC--Hercules Monochrome Card
- *.DOC--instructions on program use
- *.PAS--source code, in Turbo Pascal

ALIGN

Dan Davison
Dept of Biochemical and Biophysical Sciences
University of Houston
University Park
4800 Calhoun
Houston, TX 77004

Goad.Davison@BIONET-20.arpa

Keith Thompson
Biology Dept.
Brookhaven Nat'l Lab
Long Island, NY

The ALIGN.DOC file provides a detailed description of how the program compares sequences the user has submitted. The documentation includes a step-by-step explanation of the procedures involved, and clarifies program parameters. Several types of output are available. The program has the capacity to print any or all of the following:

1. imput data (amino acid or nucleotide sequences typed in by user)
2. table of matches--which show actual start and stop positions of each match found
3. a listing of matched and unmatched areas

System requirements: IBM PC compatible computer

Files Included:

- ALIGN.EXE--executable version of ALIGN
- ALIGN.DOC--documentation and explanation of program use

There are many other accessory files also on this disk.

ALP3/ALN3

Osamu Gotoh
Dept. of Biochemistry
Saitama Cancer Center Research Institute
Ina-machi Saitama 362
JAPAN

This diskette contains some implementations of the algorithm for aligning three protein or DNA sequences described in Gotoh, O. (1986) J. Theor. Biol. 121, 327-337.

System requirements: IBM PC compatible computer

Files included:

- ALN3.EXE--program for aligning three DNA sequences
- ALP3.EXE--program for aligning three protein sequences
- MDM-1.DAT--program data file
- MAKMDM.EXE--accessory file

Test data files:

• S1.SEQ	DNA
• S2.SEQ	"
• S3.SEQ	"
P1.PEP	Protein
• P2.PEP	"
• P3.PEP	"

Also included:

- SEQFORM.DOC --describes the sequence file format to use when running ALN3/ALP3.

PLASMID PAINT

Joe Lipsick
Dept. of Pathology, M-012
U C San Diego
La Jolla, CA 92013

Jlipsick@BIONET-20.arpa

PLASMID PAINT, a program written in Microsoft QuickBASIC 2.0, allows one to draw plasmids.

System requirements: IBM PC compatible computer with a CGA-compatible adapter and screen

Files included:

- PLASMIDC.EXE--executable file
- PAINT.BAT--short batch file which loads the DOS GRAPHICS command and then runs PLASMIDC.EXE.
- README.DOC--explanation of program use

OLIGO MUTANT MAKER

Kevin Beadles
1044 1/2 Shrader St.
San Francisco, CA 94117

(415)759-0148 (Kevin will call you back collect if he must return your call.

OligoMutantMaker simplifies the designing and screening of oligonucleotide-directed single amino acid substitution experiments by searching for nucleotide sequences which introduce a restriction endonuclease recognition sequence into the codon substitution site of the mutant.

System requirements: IBM PC compatible computer.

Files included:

- README.DOC--Documentation file
- BWMUTANT.COM--Executable file for monochrome monitors.
- CMUTANT.COM--executable file for color monitors.
- CUTTERS.DAT--Binary database of enzyme cleavage and availability data.
- CODONID.DAT--Standard genetic code.
- [ENZYME.TXT]--This file is created every time the program is run. It contains a copy of the results of the last analysis.

V. BIOSCI Bulletin Board Network Information

A copy of the information sheet which is mailed electronically to interested parties is included in this section.

BIOSCI BULLETINS

The following is a list of bboards available for distribution to sites on the ARPANET/Internet, BITNET, EARN, NETNORTH, and JANET. For each of these bboards a list of BITNET name abbreviations and analogous USENET newsgroup names are listed below. Finally, we provide a list of the various sites (nodes) that distribute the bboards and the address format for posting messages. Note that messages posted at any node are automatically redistributed to all other nodes on the BIOSCI network and subsequently to their readers.

BBOARD NAME	TOPIC
AGEING	Scientific Interest Group
BIONEWS	General announcements
BIOTECH	Biotechnology issues
BIO-CONVERSION	Scientific Interest Group
BIO-MATRIX	Applications of computers to biological databases
CONTRIBUTED-SOFTWARE	Information on molecular biology programs contributed to the public domain
EMBL-DATABANK	Messages to and from the EMBL database staff
EMPLOYMENT	Job opportunities
GENBANK-BB	Messages to and from the GenBank database staff
GENE-EXPRESSION	Scientific Interest Group
GENOMIC-ORGANIZATION	Scientific Interest Group
METHODS-AND-REAGENTS	Requests for information and lab reagents
MOLECULAR-EVOLUTION	Scientific Interest Group
ONCOGENES	Scientific Interest Group
PC-COMMUNICATIONS	Information on communications software
PC-SOFTWARE	Information on PC-software for scientists
PIR	Messages to and from the PIR database staff
PLANT-MOLECULAR-BIOLOGY	Scientific Interest Group
PROTEIN-ANALYSIS	Scientific Interest Group
RESEARCH-NEWS	Research news of interest to the community
SCIENCE-RESOURCES	Information about funding agencies, etc.
SWISS-PROT	Messages to and from the SWISS-PROT database staff
YEAST-GENETICS	Scientific Interest Group

BITNET abbreviations (<= 8 characters) for each bboard have been established:

BBOARD NAME	BITNET/EARN Name
AGEING	AGEING
BIONEWS	BIONEWS
BIOTECH	BIOTECH
BIO-CONVERSION	BIO-CONV
BIO-MATRIX	BIOMATRX
CONTRIBUTED-SOFTWARE	SOFT-CON
EMBL-DATABANK	EMBL-DB
EMPLOYMENT	BIOJOBS
GENBANK-BB	GENBANKB
GENE-EXPRESSION	GENE-EXP
GENOMIC-ORGANIZATION	GENE-ORG
METHODS-AND-REAGENTS	METHODS
MOLECULAR-EVOLUTION	MOL-EVOL
ONCOGENES	ONCOGENE
PC-COMMUNICATIONS	SOFT-COM
PC-SOFTWARE	SOFT-PC
PIR	PIR-BB
PLANT-MOLECULAR-BIOLOGY	PLANT

PROTEIN-ANALYSIS	PROTEINS
RESEARCH-NEWS	RESEARCH
SCIENCE-RESOURCES	SCI-RES
SWISS-PROT	SWISSPRT
YEAST-GENETICS	YEAST

Note: For the database bboards addresses such as PIR, EMBL, etc., were avoided since these may conflict with other BITNET addresses at the EMBL, etc., if they join the distribution scheme.

Equivalences of the Unix USENET newsgroup names to the ARPANET mailing list names follow:

bionet.general	BIONEWS
bionet.jobs	EMPLOYMENT
bionet.technology.general	BIOTECH
bionet.technology.conversion	BIO-CONVERSION
bionet.molbio.news	RESEARCH-NEWS
bionet.molbio.ageing	AGEING
bionet.molbio.bio-matrix	BIO-MATRIX
bionet.molbio.methds-reagnts	METHODS-AND-REAGANTS
bionet.molbio.genbank	GENBANK-BB
bionet.molbio.embldatabank	EMBL-DATABANK
bionet.molbio.pir	PIR
bionet.molbio.evolution	MOLECULAR-EVOLUTION
bionet.molbio.gene-express	GENE-EXPRESSION
bionet.molbio.gene-org	GENOMIC-ORGANIZATION
bionet.molbio.oncogenes	ONCOGENES
bionet.molbio.plant	PLANT-GENETICS
bionet.molbio.proteins	PROTEIN-ANALYSIS
bionet.molbio.swiss-prot	SWISS-PROT
bionet.molbio.yeast	YEAST-GENETICS
bionet.sci-resources	SCIENCE-RESOURCES
bionet.software.pc	PC-SOFTWARE
bionet.software.pc.comm	PC-COMMUNICATION
bionet.software.contrib	CONTRIBUTED-SOFTWARE

BIOSCI Nodes

Information about the BIOSCI bboard network can be obtained by mailing to the address

biosci@xxxx

where xxxx can be any of the following node addresses, e.g., biosci@uk.ac.daresbury in the United Kingdom. Interested parties outside of Europe and North America should contact whichever node is most convenient. Messages can be posted directly to bboards at any of these nodes by using the address format

bboard@xxxx

where bboard is a name from the lists above and xxxx is from the node list below, e.g., bionews@umdc.bitnet Note that nodes listed as (Internet) sites utilize the long bboard names as indicated in the first list above and nodes listed as BITNET, EARN, or JANET use the BITNET abbreviated bboard names.

Europe		
Sweden	bmc.uu.se	(Internet)
UK	uk.ac.daresbury	(JANET)
Ireland	irlearn.bitnet	(BITNET/EARN)
Ireland	irlearn.ucd.ie	(Internet, but uses BITNET bboard names)

North America
net.bio.net (Internet)
bionet-20.bio.net (Internet)
umdc.bitnet (coming soon, BITNET)

VI. BIONET Training Schedules

The schedules for the five Mountain View Trainings are included.

Training Schedule

March 17 & 18, 1988

Thursday, March 17

Time	Topic	Instructor
9:00- 9:45	Introduction	Nancy Bigham
9:45-10:15	Sequence Entry	
10:15-10:30	Break	
10:30-12:00	Sequencing Project Management	Vicki Johncox
12:00- 1:00	Lunch	
1:00- 1:20	Cloner Demonstration	Constance Gertsch
1:30- 2:00	Sequence Alignment	David Kristofferson
2:00- 3:15	DNA Sequence Analysis	
3:15- 3:30	Break	
3:30- 4:45	Peptide Sequence Analysis	Spencer Yeh

Friday, March 18

Time	Topic	Instructor
9:00- 10:30	Database Structure and Simple Searches	Constance Gertsch
10:30-10:45	Break	
10:30-12:00	Finding Sequences in the Databases	
12:00- 1:00	Lunch	
1:00- 3:00	Sequence Alignment	Sunil Maulik
3:00- 3:15	Break	
3:15- 4:00	Electronic mail and bulletin boards	Dave Kristofferson

An Introduction to the IntelliGenetics Suite
Course Schedule
May 19 and 20, 1988

Thursday, May 19

8:30	9:00	Introduction Overview of the IntelliGenetics Programs	Nancy Bigham
9:00	9:45	Sequence Data Entry and Editing	
9:45	10:30	System commands	
10:30	10:45	Break	
10:45	12:00	Sequencing Gel Management	Vicki Johncox
12:00	1:00	Lunch	
1:00	2:15	Nucleic Acid Sequence Analysis	Trish Benton
2:15	2:30	Break	
2:30	3:45	Peptide Sequence Analysis	Spencer Yeh
3:45	5:00	Electronic Mail and Bulletin Boards File Transfer	Dave Kristofferson

Friday, May 20

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Beth Swank
10:30	10:45	Break	
10:45	12:00	Collecting related sequences, searching for patterns	
12:00	1:00	Lunch	
1:00	3:15	Sequence Similarity Searches	Sunil Maulik
3:15	3:30	Break	
3:30	4:30	Review and Questions	Nancy Bigham

An Introduction to the IntelliGenetics Suite
Course Schedule
July 14 and 14 1988

Thursday, July 19

8:30	8:45	Introduction Overview of IG Programs	Vickie Johncox
8:45	9:00	BIONET Resource	Dave Kristofferson
9:00	9:45	Operating System, File Structure	Vickie Johncox
9:45	10:30	Electronic Mail	Kathy Berg
10:30	10:45	Break	
10:45	12:00	Sequence Entry	Vickie Johncox
12:00	1:00	Lunch	
1:00	2:00	Sequencing Gel Management	Spencer Yeh
2:00	3:15	NA and Peptide Sequence Analysis	
3:15	3:30	Break	
3:30	5:00	Problems	

Friday, July 20

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Vickie Johncox
10:30	10:45	Break	
10:45	12:00	Collecting Sequences, Pattern Searching	
12:00	1:00	Lunch	
1:00	3:15	Sequence Similarity Searches	Sunil Maulik
3:15	3:30	Break	
3:30	4:30	Review and Questions	Vickie Johncox

An Introduction to the IntelliGenetics Suite
Course Schedule
September 15 and 16 1988

Thursday, September 15

8:30	8:45	Introduction	Vickie Johncox
8:45	9:00	Introduction to BIONET	Dave Kristofferson
9:00	9:30	Operating System, File Structure	Trish Benton
9:30	10:00	Electronic Mail	Kathy Berg
10:00	10:15	Organization of IntelliGenetics	Murray Summers
10:15	10:30	Break	
10:30	11:15	Sequence Data Entry and Editing	Karen Davis
11:15	12:00	Sequencing Gel Management	Trish Benton
12:00	1:00	Lunch	
1:00	3:15	NA and Peptide Sequence Analysis	Karen Davis
3:15	3:30	Break	
3:30	5:00	Problems	

Friday, September 16

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Trish Benton
10:30	10:45	Break	
10:45	12:00	Collecting Sequences, Pattern Searching	
12:00	1:00	Lunch	
1:00	3:15	Sequence Similarity Searches	Sunil Maulik
3:15	3:30	Break	
3:30	4:00	Problems	
4:00	4:15	Review and Questions	Vickie Johncox

An Introduction to the IntelliGenetics Suite and BIONET
Course Schedule
November 17 and 18, 1988

Thursday, November 17

8:30	9:00	Introduction	Vickie Johncox
9:00	9:15	Introduction to BIONET	Dave Kristofferson
9:15	9:45	Operating System, File Structure	Vickie Johncox
9:45	10:15	Electronic Mail	
10:15	10:30	BREAK	
10:30	11:15	Sequence Data Entry and Editing	
11:15	12:00	Sequencing Gel Management ,	
12:00	1:00	LUNCH	
1:00	1:30	Problems	
1:30	3:30	NA and Peptide Sequence Analysis	Karen Davis
3:30	3:45	BREAK	
3:45	5:00	Problems	
5:00	6:30	BREAK	
6:30	9:00	BIONET-specific topics	BIONET staff

An Introduction to the IntelliGenetics Suite and BIONET
Course Schedule
November 17 and 18, 1988

Friday, November 18

8:00	9:00	Open Classroom	Consultants
9:00	10:30	Database Structure and Sequence Retrieval	Karen Davis
10:30	10:45	BREAK	
10:45	12:00	Collecting Sequences, Pattern Searching	Karen Davis
12:00	1:00	LUNCH	
1:00	3:00	Sequence Similarity Searches	Vickie Johncox
3:00	3:15	BREAK	
3:15	3:45	Problems	
3:45	4:00	Review and Questions	Vickie Johncox
4:00	5:00	BIONET-specific topics	BIONET staff

Introduction to BIONET
Thursday Evening Schedule
November 17, 1988

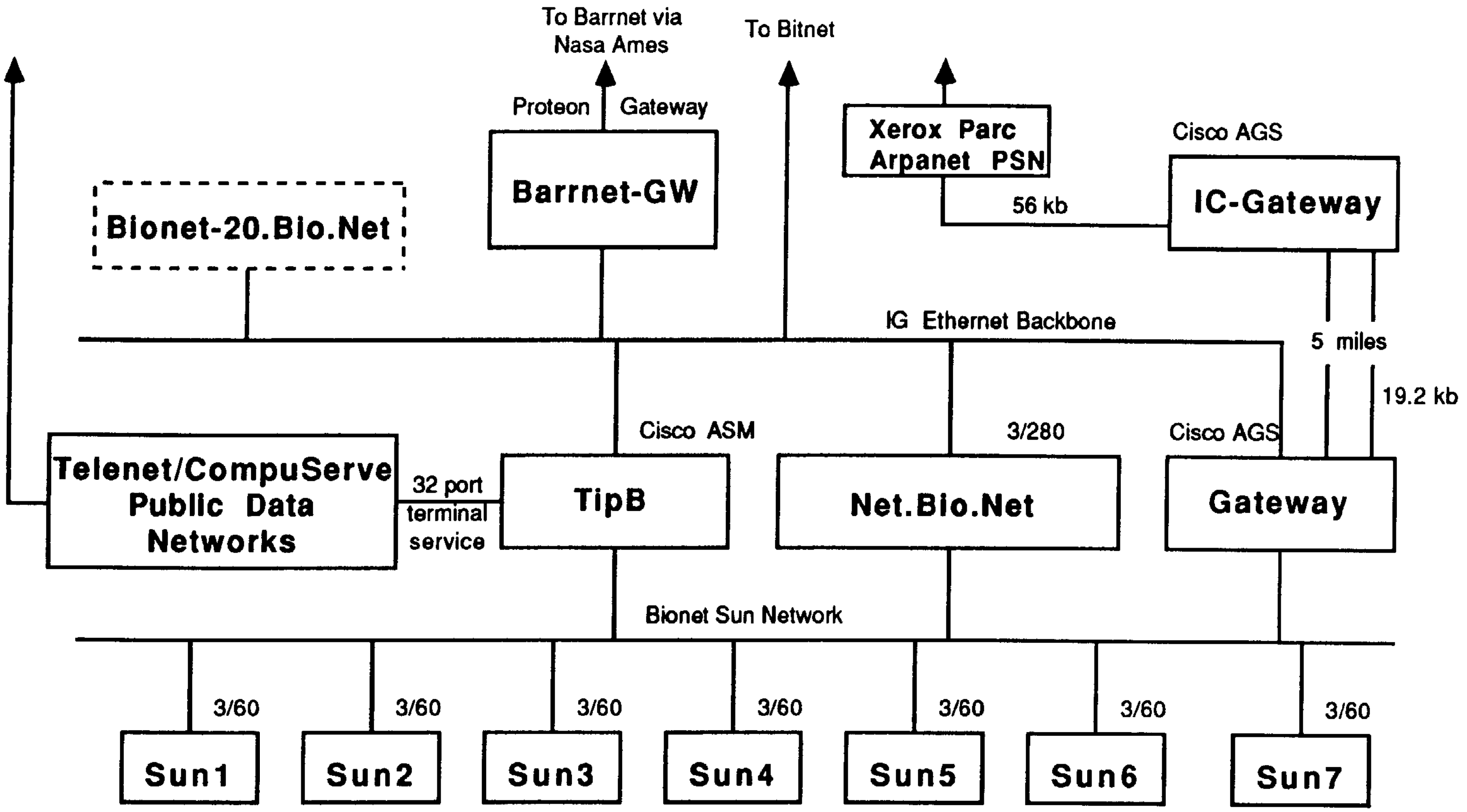
6:30	7:15	Connecting to BIONET, File Transfers	David Kristofferson
7:15	7:45	TOPS-20 Operating System, EMACS Text Editor	Karen Davis
7:45	8:00	BREAK	
8:00	8:30	Bulletin Boards, E-mail Addresses, Help Me	Karen Davis
8:30	9:00	Other BIONET Software	Spencer Yeh

Introduction to BIONET
Friday Afternoon Schedule
November 18, 1988

4:00	5:00	FASTA-MAIL	Spencer Yeh

VII. BIONET Computer Facilities

A diagram of the BIONET computer facilities follows on the next page.



Bionet Central Computing Resource

VIII. Testimonials

Copies of two recently received testimonial letters are included in this section.

HARVARD MEDICAL SCHOOL

DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY

Tel. (617) 732- 2046
Fax: 738-0516
Internet: Hirsh@BIONET-20.bio.net

25 Shattuck Street
Boston, Massachusetts 02115

Jay Hirsh
Assoc. Professor
260 Longwood Ave.

October 21, 1988

Re: Statement in support of NIH grant to the BIONET computer facility.

To whom it may concern:

We have been members of the BIONET computer facility for approximately 3 years. We have found this facility to be of great importance to our research. We make extensive use of the homology searching and sequence analysis routines, and are just beginnning to take full advantage of the database searching routines. These analyses have uncovered a number of promoter elements of the Drosophila dopa decarboxylase gene (*Ddc*) that are conserved through evolution and are functionally important ((Scholnick et al (1986) *Science* **234**, 998-1002; Bray & Hirsh (1986) EMBO J., 5,2305-2312; Bray et al (1988) EMBO J.,7,177-188.). We enclose a copy of this last reference. Even though it did not utilize directly the Bionet resource, this manuscript shows directly that one of the initially identified conserved elements is a CNS-specific regulatory element.

We have recently extended this analysis to a cell-specfic enhancer of *Ddc* (Johnson, McCormick, Bray & Hirsh, in preparation), an ~800 bp segment that contains a number of functional elements. This approach is also proving to be valid in this region, in that a number of conserved elements are showing interesting in vivo functions in regulating different aspects of the neuronal pattern of expression of *Ddc*.

Within the past year, we have also begun to make extensive use of the Bionet E-mail facility. We now routinely communicate with European and American colleagues using this system. Given the cost of international phone calls, our use of this facility alone would pretty much justify the yearly bionet fee.

We continue to use this facility, even though there is presently a local computer facility claiming to offer comparable services at a comparable cost. This decision is due to the superb level of support that we have derived from the Bionet staff: Questions regarding operation of the system are answered expertly and promptly. The programs and program manuals have been continually evolving, such that the programs are now accessible to even novice users in the lab. When I have made queries-such as wondering whether database searches couldn't be done without hanging on the phone for hours- I have been surprised and amazed that the staff have gone to what appear to be major efforts to implement such searching programs. Our experiences with the aformentioned local computer facility, where the level of staff support was almost nill, have shown the value of such services.

In summary, I hope that NIH will continue to support the Bionet resource. This program has attracted a dedicated and skilled staff that provides services and support that we have not been able to find elsewhere. An interruption in the funding of this program would certainly cause disruption to a large number of users who are highly dependent on these people and this system.

Sincerely,

Jay Hirsh

THE JOHNS HOPKINS UNIVERSITY

BALTIMORE, MARYLAND 21218

DEPARTMENT OF BIOLOGY

November 28, 1988

Dr. David Kristofferson
BIONET
c/o Intelligenetics
700 E. El Camino Road
Mountain View, CA 94040

Dear Dave,

I am addressing this letter to you personally because I dislike the "To Whom It May Concern" format - it makes me feel as though I am writing to an answering machine. Please feel free however to show the contents to anyone.

There are several statements that I would like to make about BIONET, its services, its staff and the Intelligenetics suite of DNA and protein sequence analysis software that they offer.

1. BIONET offers a level of support for the researcher that cannot be duplicated by many universities or research centers.

While DNA/protein sequence analysis software can be put on the university or medical school mainframe computer or run on personal computers, the problem has always been using it: training people in fundamentals (try to find understandable documentation on some of these programs) and then teaching them the significance of the variables involved and the limitations of the algorithms used. BIONET has made a considerable effort to address these problems. They offer affordable training at Mountain View, California and are willing to come to you if they have staff and time. They have made a real attempt to write complete and understandable and useful documentation for the system and the programs. Most important is their online help and on-call systems experts who answer immediate questions about programs and data. It is difficult for a systems operator at a university computer center to be well-versed on all the software running on the machine. Even if they are, they need the time to explain the programs. These overworked systems operators must rely on local experts or bright students to help people with questions about individual programs. Molecular biologists need easily accessible, experienced professionals to answer their questions. BIONET offers those professionals.

2. The BIONET staff are highly competent and extremely cooperative. I have nothing but praise for the personnel at BIONET and the service and support they offer. Over the last year, I have received advice (good advice, by the way), had problems solved, and talked about on-site training. Everyone who dealt with me was knowledgeable, professional and very helpful. I really appreciate having these people on call to answer my questions, which they do with remarkable rapidity and skill.

3. The systems analysts and programmers at BIONET listen to the users. Several times I have had questions about the running of a program or suggestions about format or documentation and the people at BIONET always listened and corrected the program's problem or explained to me my problem. They are always improving the system and it is wonderful never to be stuck with a "quirk" in the program.

4. Access to BIONET is very easy for the computer novice. As the computer "expert" in Dr. Roseman's laboratory, I find that people are very cautious about accessing the mainframe computer on campus. Mistakes cost money and since the system is not dedicated to one suite of software, the interface has to be more complicated. Most people don't want to learn DOS for the PC let alone a whole new operating system for the Vax or the IBM. BIONET allows the novice easy access while retaining all the system commands for the more expert or more daring among us. People in my laboratory are connecting to the BIONET system with ease and learning as they work. This is an immense time-saver for me since the online help keeps them from calling my name every five minutes and distracting me from my own research.

5. BIONET provides communication lines between scientists that are easily used and therefore frequently used. For a while BIONET was the only large scale bulletin board system for molecular biologists. Now more and more communications lines are open, but the problem is again getting people to use them. Some of them require Bitnet or Arpanet, etc. and that requires people to access their mainframe computers and learn how to run electronic mail systems that are not as user-friendly as the BIONET system. BIONET is now acting as a gateway to a number of these systems and I for one, am grateful that I can hear the news, or send the news without running one more program.

I am very pleased with BIONET and the Intelligenetics programs. They have enabled my colleagues and me to further our work in molecular biology by providing the sequence analysis software in a form that we could learn and understand. I believe that BIONET provides necessary and unique services for the research scientist.

Very truly yours,
Donna K. Fox
Donna K. Fox, Ph.D.
Associate Research Scientist